Optimal replacement in the proportional hazards model and its applications in a product-service system

by

Xiang Wu

A dissertation submitted to the graduate faculty

In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Sarah M. Ryan, Major Professor
William Q. Meeker
James D. McCalley
Sigurdur Olafsson
Lizhi Wang

Iowa State University

Ames, Iowa

2012

DEDICATION

I would like to dedicate this thesis to my wife, Bei Huang; without whose love and support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deepest gratitude to Dr. Sarah M. Ryan for her insightful guidance, constructive suggestions, patient instructions and particularly constant encouragement during my graduate study. She is the best advisor that I could ever ask for. I am also very grateful to my committee members, Dr. Meeker, Dr. McCalley, Dr. Olafsson and Dr. Wang for their invaluable advice.

The support from the National Science Foundation grant CNS-0540293 is also gratefully acknowledged.

ABSTRACT

Condition-based maintenance is rapidly gaining favor as a way to prevent the failures of capital-intensive assets and to maintain them in good operating condition with minimum cost. A valuable and increasingly prevalent way to incorporate condition information into risk estimation is by the proportional hazards model (PHM), which explicitly includes both the age and the condition information in the calculation of the hazard function. This dissertation consists of three papers, in which the optimal replacement policies for systems whose deterioration process follows the PHM are developed under different settings; and a joint optimization of the asset and inventory management problem in the context of a product-service system is considered.

In the first paper, a continuous time Markov covariate process is assumed to describe the condition of a system that is under periodic monitoring. Although the form of an optimal replacement policy for such a system in the PHM was developed previously, an approximation of the Markov process as constant within inspection intervals led to a counter-intuitive result that less frequent monitoring could yield a replacement policy with lower average cost. Accounting for possible state transitions between inspection epochs removes the approximation and eliminates the cost anomaly. A new recursive procedure to obtain the parameters of the optimal replacement policy is presented. By comparing the replacement and monitoring costs of different monitoring scheme, the value of condition information is evaluated.

In the second paper, the optimal replacement policy for systems in the PHM with semi-Markovian covariate process and continuous monitoring is developed. Numerical examples and sensitivity analysis provide some insights about the suitability of a Markov approximation and the impact of the variations in the input parameters on the cost.

In applying the optimal replacement policies to a product-service system, where the producers provide the use of the products to customers while retaining ownership, the coupling between the decision making for preventive replacement and the decision making

for inventory management is evident. In the third paper, an integrated model is proposed for the preventive maintenance of a fleet of products and the inventory management of a hybrid manufacturing-remanufacturing system in the context of a product-service system. A joint optimization technique is developed to obtain the optimal parameters for the operational policy of the integrated model to minimize the long run average cost per unit time. In addition, the effect of the assumption that the replaced products are not sorted is evaluated.

CHAPTER 1 GENERAL INTRODUCTION

1.1 Motivation

Critical infrastructures depend on equipment and systems that deteriorate with age and are subject to failure. Because abrupt failures of capital-intensive physical assets such as high-voltage power transformers and heavy mining equipment may cause immense economic loss, preventive maintenance is essential.

Optimal maintenance policies for deteriorating systems have been extensively studied for decades, and the recent research effort has been focused on condition-based maintenance (CBM). Compared to classical age-based preventive maintenance, CBM improves the decision-making process greatly by exploiting available information about the system's operating conditions, such as use rate, temperature, humidity, vibration levels, or the amount of metal particles in the lubricant etc., in addition to the age information (Banjevic et al., 2001).

CBM relies heavily on condition monitoring technology. Increasingly, condition monitoring technology is gaining favor as a way to diagnose the health status, detect abnormal conditions and prevent catastrophic failure of valuable assets. Generally, there are two types of condition monitoring: periodic monitoring and continuous monitoring. The most rudimentary form of condition monitoring is periodic visual inspection by experienced operators to detect failure indicators such as cracking, leaking, corrosion, etc. More advanced periodic monitoring can be done by personnel through handheld data collectors and analyzers to collect information from oil analysis, vibration analysis, ultrasound analysis etc.

With the advance of information and communication technology, remote and continuous monitoring of the condition information becomes accessible to decision makers. Today it is possible to install sensors and smart chips in a product to measure and record use rate/environmental data over the life of the product, and those data can be returned in real time to a central location to aid CBM decision making (Hong, 2009). This type of monitoring

is appealing particularly when distance or environmental conditions make regular inspections difficult.

Condition monitoring may require substantial initial investment. Implementing CBM requires installation of instruments such as thermal sensors, debris detectors, dissolved gas analyzers or vibration monitors plus the information and communication devices to collect and transmit the condition data. The cost of sufficient instruments can be quite high, especially on equipment that is already installed. Therefore it is of vital importance to justify the value of condition monitoring before adding it to all equipment.

In the first paper, two related challenging questions are addressed: 1) how to make best use of the condition information; and 2) whether the investment in condition monitoring technology is worthwhile.

CBM models differ according to the approaches for utilizing the condition information to model the system's lifetime. Many researchers assume that the system failure process can be described adequately by a multi-state deteriorating model derived from condition information, and extensive research has been done with Markov and semi-Markov decision models (Mine and Kawai, 1975, Mine and Kawai, 1982, Lam and Yeh, 1994, Chen and Trivedi, 2005). In this research, the proportional hazards (PH) model (Cox and Oakes, 1984) is adopted to incorporate condition information into system risk estimation. The condition information can be considered as a vector of covariates, each representing a certain measurement. The PH model combines a baseline hazard function which accounts for the aging degradation with a link function that takes the covariates into account to improve the prediction of failure. Compared to the Markovian deteriorating models, the PH model provides explicit expressions of the hazard function and the failure probability, which is more convenient to use and easier to calibrate from statistical point of view, and therefore is more accurate.

Using a continuous time Markov chain to describe the evolution of the system condition, the form of an optimal replacement policy for systems that follow the PH model was developed by Makis and Jardine (1992). However, their approximation of the Markov process as constant within inspection intervals led to a counter-intuitive result that less frequent monitoring could yield a replacement policy with lower average cost. We explicitly

account for possible state transitions between inspection epochs to remove the approximation and eliminate the cost anomaly, and present a new recursive procedure to compute the optimal replacement age in each of the operating states and the optimal average cost for periodic monitoring. This allows an accurate comparison of monitoring at discrete intervals of different lengths against continuous monitoring (approximated as the interval vanishes) or no monitoring.

We compare the average cost per unit time for monitoring and replacement under three monitoring schemes: no monitoring which corresponds to age-based replacement, periodic monitoring at various intervals, and continuous monitoring. We characterize the cost environments in which investment in condition monitoring equipment is justified.

The second paper generalizes the CBM model in the first paper in two aspects. First the stochastic process characterizing the condition information is extended from a Markov process to a semi-Markov process, which allows arbitrary sojourn time distributions between transitions among the covariate states. Second the optimal replacement problem for systems in the PH model under continuous monitoring is investigated and a procedure is developed to obtain the optimal parameters and costs when the condition information is continuously available. Those generalizations endow our method with more capability and flexibility to model real world situations. In addition, sensitivity analysis is performed on a specific instance to demonstrate how the variations in the input parameters would affect the long-run average cost.

A product-service system (PSS) is a strategy in which producers provide the use as well as the maintenance of products while retaining ownership. Prospective customers who become the clients pay fees for receiving the services or functions of products rather than purchasing them, and so are free of the risk, responsibility and cost burdens which are commonly associated with ownership. Since the introduction of this attractive concept in 1999 (Goedkoop et al., 1999, White et al., 1999), a diverse range of PSS examples in the literature have demonstrated its economic success as well as its significant environmental benefits and social gains (Luiten et al., 2001, Manzini et al., 2001, Baines and Lightfoot, 2007).

In PSS, a provider must maintain its products continuously in working order while they are dispersed among client firms. Because it retains ownership and control over the products, it can reuse and remanufacture them extensively. These practices motivate the use of condition monitoring to increase visibility of the product's condition and environment while in use. In the attempt to apply CBM to the product-service system, the coupling between the decision making for preventive maintenance and the decision making for inventory management is evident.

In particular, for the service paradigm to be viable from the provider's perspective, the fee for service must allow for a profit margin over the cost of providing the service. The cost of service provision depends largely on the ability to manage and maintain products effectively in a closed-loop system. Unlike the common closed-loop supply chain for sold products, a distinct feature of the closed-loop supply chain in PSS is that the demands are driven by replacement of products in service and/or a capacity expansion requirement, and the returns are essentially generated by out-of-service products, replaced either preventively or due to failure. In other words, the demands and returns are controllable by the provider via replacement decisions, and the cost of replacement is affected by the inventory management decisions. Therefore, the replacement decisions are closely coupled with the inventory management decisions of this closed-loop supply chain. This coupling makes the decision making under PSS significantly more complicated than that under traditional product sales.

In the third paper, we present an integrated model which takes into account both the maintenance decisions and the inventory management decisions in the context of a product-service system to minimize the total cost per unit time. For maintenance, we consider the condition-based replacement policy presented in the second paper. For inventory management, a continuous review base stock policy is adopted due to its easy implementation and proven effectiveness in practice. Identifying and formulating the couplings between them, we develop an optimization technique to obtain the optimal parameters for the two policies simultaneously in the integrated model. In addition, the effect of the assumption that the replaced products have no quality difference is evaluated.

1.2 Dissertation Organization

This dissertation consists of three main chapters, preceded by this general introduction and followed by a general conclusion. Each of those main chapters is a journal article, with the first two published and the third under review. Chapter 2 assesses the value of condition information for optimal replacement in the proportional hazards model with continuous degradation. Chapter 3 investigates the optimal replacement in the proportional hazards model with semi-Markovian covariate process and continuous monitoring. Chapter 4 studies the joint optimization of the asset and inventory management in the context of product-service system. Chapter 5 concludes.

References

Baines, T. S. and Lightfoot, H. W. (2007). State-of-the-art in product-service systems. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture,* 221:1543–1552.

Banjevic, D., Jardine, A. K. S., Makis, V., and Ennis, M. (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR*, 39:32–50.

Chen, D. and Trivedi, K. S. (2005). Optimization for condition-based maintenance with semi-Markov decision process. *Reliability Engineering & System Safety*, 90(1):25–29.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Goedkoop, M., van Halen, C., and te Riele, H. (1999). Product service-systems, ecological and economic basics. Report for Dutch ministries of environment (VROM) and economic affairs (EZ). http://www.pre.nl/download/ProductService.zip.

Hong, Y. (2009). *Reliability prediction based on complicated data and dynamic data*. PhD thesis, Iowa State University.

Lam, C. T. and Yeh, R. H. (1994). Optimal replacement policies for multi-state deteriorating systems. *Naval Research Logistics*, 41(33):303–315.

Luiten, H., Knot, M., and van der Horst, T. (2001). Sustainable product-service-systems: the kathalys method. In *Proceedings of the Second International Symposium on Environmentally conscious design and inverse manufacturing*, pages 190–197.

Makis, V. and Jardine, A. K. S. (1992). Optimal replacement in the proportional hazards model. *INFOR*, 30(1):172–183.

Manzini, E., Vezzoli, C., and Clark, G. (2001). Product service-systems: using an existing concept as a new approach to sustainability. *Journal of Design Research*, 1(2).

Mine, H. and Kawai, H. (1975). An optimal inspection and replacement policy. *IEEE Transactions on Reliability*, 24:305–309.

Mine, H. and Kawai, H. (1982). An optimal inspection and replacement policy of a deteriorating system. *Journal of Operations Research Society of Japan*, 25:1–15.

White, A., Stoughton, M., and Feng, L. (1999). *Servicizing: The Quiet Transition to Extended Producer Responsibility*. Tellus Institute, Boston.

CHAPTER 2 VALUE OF CONDITION MONITORING FOR OPTIMAL REPLACEMENT IN THE PROPORTIONAL HAZARDS MODEL WITH CONTINUOUS DEGRADATION

A paper published in *IIE Transactions*[1]

Xiang Wu and Sarah M. Ryan

Abstract

We investigate the value of perfect monitoring information for optimal replacement of deteriorating systems in the proportional hazards model (PHM). A continuous time Markov chain describes the condition of the system. Although the form of an optimal replacement policy for system under periodic monitoring in the PHM was developed previously, an approximation of the Markov process as constant within inspection intervals led to a counter-intuitive result that less frequent monitoring could yield a replacement policy with lower average cost. We explicitly account for possible state transitions between inspection epochs to remove the approximation and eliminate the cost anomaly. However, the mathematical evaluation becomes significantly more complicated. To overcome this difficulty, we present a new recursive procedure to obtain the parameters of the optimal replacement policy and the optimal average cost. A numerical example is provided to illustrate the computational procedure and the value of condition monitoring. By taking the monitoring cost into consideration, we observe the relationships between the unit cost of periodic monitoring and the upfront cost of continuous monitoring under which the continuous, periodic or no monitoring scheme is optimal.

Keyword: optimal replacement; proportional hazards model; continuous time Markov chain; value of condition monitoring.

---

[1] Appeared in *IIE Transactions*, 2010, 42, 553-563

2.1 Introduction

Critical infrastructures depend on equipment and systems that deteriorate with age and are subject to failure. Because abrupt failures of assets such as high-voltage power transformers and heavy mining equipment may cause immense economic loss, preventive maintenance is essential. Some of these assets or their electronic components are difficult and/or exorbitantly expensive to repair, and the need for continuous service precludes shutting down the dependent systems while on-site maintenance or repairs are done. In this paper, we consider replacement as the only maintenance option.

Optimal replacement policies for deteriorating systems have been studied for decades (Aven and Bergman, 1986; Lam and Yeh, 1994b), and the recent research effort has been focused on the problem of optimal replacement when some concomitant (condition) information about the system, such as temperature, humidity, vibration levels, or the amount of metal particles in the lubricant, is available. Remote monitoring of condition information is appealing particularly when distance or environmental conditions make regular inspections difficult. Condition monitoring sensors along with information and communication technology increase the visibility of the system's condition and environment while in use. Condition-based maintenance policies, such as those in Banjevic et al. (2001), Makis and Jiang (2003), Dieulle et al. (2003) and Ghasemi et al. (2007), exploit such information to determine when to preventively replace the system. Presumably, policies derived from more frequent observations of condition information have lower cost than those based on less frequent or no observations. The reduction in expected cost provided by frequent monitoring can be used to assess the value of the technology that enables the monitoring.

Condition monitoring may require substantial initial investment in hardware and software installation, in contrast to traditional monitoring which typically incurs a cost associated with each observation. Taking this latter cost into consideration for systems under sequential or periodic monitoring, the optimal monitoring interval is usually determined by searching the possible parameter space within each step of a policy iteration algorithm, such as those in Yeh (1997) and Chiang and Yuan (2001). Continuous monitoring has been studied more

recently (Liao et al., 2006). Comparison of periodic and continuous monitoring for a two-state system has been considered by Rosenblatt and Lee (1986). A more general comparative study of sequential and continuous monitoring strategies for a multistate model was presented by Lam and Yeh (1994a) for a deteriorating Markovian system; however, they did not include any cost for continuous monitoring.

The concomitant information may be described by a stochastic process, which most frequently appears in the literature as a semi-Markov or Markov process. Models of the system's failure probability differ according to their approaches for utilizing the condition information. Many researchers assume that the failure process of the system can be described adequately by a multi-state deteriorating Markov or semi-Markov process that leads to failure, and extensive research has been done with such models. For example, Chiang and Yuan (2001) proposed a state-dependent maintenance policy for a Markovian deteriorating system and they showed that many policies presented earlier were special cases of their proposed policy. Bloch-Mercier (2002) studied the preventive maintenance policy for a Markovian deteriorating system when a sequential checking procedure is applied. A dynamic preventive maintenance policy for a multi-state deteriorating system was developed by Chen et al. (2003).

For many applications, it is most natural to model failures as dependent on system age in addition to some deterioration process. One way to account for these combined effects is to use the proportional hazards model (PHM), which explicitly includes both the age and the condition information in the hazard function (Makis and Jardine, 1992; Banjevic et al., 2001). Makis and Jardine (1992) derived an optimal replacement policy for systems in the PHM with a continuous time Markov chain and periodic monitoring, and presented recursive methods to compute the optimal policy parameters. Banjevic et al. (2001) extended Makis and Jardine's model by relaxing the monotonicity assumption of the hazard function and they developed methods for estimating model parameters as well. However, the computations in both papers relied on approximating the concomitant Markov chain as unchanging between inspection epochs. Ghasemi et al. (2007) also used the PHM to characterize the system failure process and, under the same discrete time approximation, derived an optimal replacement policy when the condition information of the system is only partially observed.

In this paper, we compare the average cost per unit time of monitoring, replacement and failure under three monitoring schemes: no monitoring which corresponds to age-based replacement, periodic monitoring at various intervals, and continuous monitoring. For periodic monitoring, we follow the model of Makis and Jardine but remove their discrete-time approximation by explicitly accounting for the possibility that the concomitant Markov chain may make transitions among its states between observation epochs. This allows an accurate comparison of monitoring at discrete intervals of different lengths against continuous monitoring (approximated as the interval vanishes) or no monitoring. Accounting for state transitions between observations introduces significant intricacies in the computation of policy parameters. These are addressed in Sections 2.3-2.5. We use conditioning to develop a new recursive procedure to obtain the parameters of the optimal replacement policy and its long-run average cost. We focus on systems with an underlying pure-birth process having an arbitrary number of states and illustrate the reasoning and computations for a three-state deterioration process in detail. In Section 2.6 we review the optimal replacement age for the no-monitoring scheme. Section 2.7 illustrates the computation of replacement policy parameters under periodic monitoring and the overall cost comparison of the three monitoring schemes in numerical examples. Based on the numerical results, we illustrate relationships between the costs of periodic or continuous monitoring under which the different monitoring schemes minimize the overall cost. Section 2.8 concludes.

2.2 Model Description

We assume that the deterioration of the system follows a continuous time process and the system can fail at any time instant. The hazard rate of the system depends both on its age and on the values of concomitant variables that reflect the current system state or the operating environment.

We use average cost per unit time to compare three schemes for monitoring and replacement decision-making. The simplest is to choose a replacement time based only on the age of the system. In this case the cost is due only to replacements and failures. The second scheme is to inspect the condition at discrete time intervals of length Δ. We assume

each inspection costs a fixed amount γ. The third is to pay an amount Γ upfront to install equipment and software that will enable continuous monitoring with no additional cost per observation. To evaluate continuous monitoring, we approximate the replacement and failure cost using periodic monitoring with $\Delta \to 0$. The goal is to determine relationships between γ and Γ under which each of these schemes minimizes the total average cost of inspection, failure and replacement per unit time, where Δ is optimized in the second approach.

Let G_1, G_2, G_3 be the average costs per unit time of the three schemes, respectively, and let g_Δ be the minimum replacement and failure cost per unit time for a periodic monitoring scheme with a fixed interval Δ. Assume r is the interest rate for continuous discounting. Then $G_1 = G_1(\tau)$ where τ is the replacement age, $G_2 = G_2(\Delta) = g_\Delta + \frac{\gamma}{\Delta}$, and $G_3 = \Gamma' + g_0$, where $g_0 = \lim_{\Delta \to 0} g_\Delta$ and $\Gamma' \equiv r\Gamma$ is found from $\Gamma = \int_0^\infty e^{-rt}\Gamma' dt$ as the equivalent average cost per unit time of Γ.

For simplicity, we consider only one concomitant variable (covariate) in this paper. We assume that the operating condition of the system, which is described by the concomitant variable, may be classified into a finite set of states, $S = \{0,1,...,n-1\}$. State 0 is the initial state of a new system. States $1,2,.,n-1$ reflect the increasingly deteriorating working condition of the system. Upon replacement, the system returns to state 0. The transition course among the states is formulated as a diagnostic stochastic process $Z = \{Z_t, t \geq 0\}$ which is a continuous time homogeneous Markov chain on state space S.

A convenient method to include both the age effect and the condition information in the hazard rate function is to employ the proportional hazards model (PHM), which has been applied successfully to engineering reliability problems in recent years (Cox and Oakes, 1984). In the PHM, the hazard rate of a system is assumed to be the product of a baseline hazard rate $h_0(t)$ dependent only on the age of the system and a positive function $\psi(\bullet)$ that depends only on the values of concomitant variables (in our case, the states of the Z process). Thus, the hazard rate of the system at time t can be expressed as

$$h(t, Z_t) \equiv h_0(t)\psi(Z_t), t \geq 0 .$$

From the above analysis, it is obvious that the key to comparing among different monitoring schemes is to obtain the optimal replacement policy and optimal replacement cost for periodic monitoring. Thus, first we assume that the Z process is under periodic monitoring with a constant cost γ per period. In other words, the states of the Z process are available only at time instants 0, Δ, 2Δ, …, where $\Delta > 0$, in a replacement cycle.

We adopt the following notation in this paper:

t: The age of the system from time of replacement.

T: The time to failure of the system.

$Z = \{Z_t, t \geq 0\}$: A continuous time Markov chain that reflects the condition of the system at age *t* with $Z_0 = 0$; in general, the effect of the operating environment on the system.

X_k: The sojourn time of the Z process in state k, $k = 0,1,...,n-2$, assumed exponentially distributed.

v_k: The hazard rate of X_k.

$h_0(t)$: The baseline hazard rate, which depends only on the age of the system.

$\psi(Z_t)$: A link function that depends on the state of the stochastic process Z.

Δ: The length of the monitoring interval.

C: The replacement cost without failure, $C > 0$.

K: The additional cost for a failure replacement, $K > 0$.

γ: The monitoring cost per period for periodic monitoring.

Γ: The one-time initial cost for continuous monitoring.

r: Interest rate for continuous discounting.

g_Δ: Minimum replacement and failure cost per unit time for monitoring interval Δ.

In addition, we introduce the following basic assumptions:

1. The system must be kept in working order at all times. Replacement is instantaneous.
2. The continuous time Markov chain Z is a pure birth process, i.e., whenever a transition occurs the state of the system always increases by one. Replacement

restarts the process at $Z_0 = 0$ and state *n*-1 is absorbing. Note that the Markov chain governs how the condition variable evolves without intervention. If maintenance actions other than replacement were considered in the model, this monotonicity assumption would be violated.

3. The baseline hazard rate, $h_0(t)$, is a non-decreasing function of the system age, that is, the system deteriorates with time.
4. The link function, $\psi(Z_t)$, is a non-decreasing function with $\psi(0) = 0$.
5. The practice of periodic monitoring influences neither the diagnostic Z process nor the system failure process.
6. Failure of the system can occur at any time. Upon failure, system replacement is executed immediately.
7. The pair (I_t, Z_t), where $I_t = 1$ if $T > t$ and 0 otherwise, is a Markov process in the following sense: For any times $0 \le s_0 < s_1 < \ldots < s_{k-1} < s < t$ and states $i_0, i_1, \ldots, i_{k-1}, i, j$,

$$P\left(T > t, Z_t = j \middle| T > s, Z_s = i, Z_{s_{k-1}} = i_{k-1}, \ldots, Z_{s_0} = i_0\right) = P\left(T > t, Z_t = j \middle| T > s, Z_s = i\right).$$

As discussed by Banjevic et al. (2001), Z_t could represent either an "external" covariate such as environmental condition or an "internal" diagnostic variable.

Under periodic monitoring, let $Z_{k\Delta}$ be the condition at time point $k\Delta$ after the most recent replacement. Although condition information is available only at integer multiples of Δ, the continuous time Markov chain Z_t may shift among its discrete values at any time. Then for $t \in [0, \Delta]$, define the expected conditional reliability function

$$\bar{R}(k, Z_{k\Delta}, t) \equiv E\left[P(T > k\Delta + t \mid T > k\Delta, Z_\Delta, ..., Z_{k\Delta})\right] = E\left[\exp\left(-\int_{k\Delta}^{k\Delta+t} h_0(s)\psi(Z_s)ds\right)\middle| Z_{k\Delta}\right] \quad (2.1)$$

This expression for the reliability function differs from the one in Makis and Jardine (1992). In the previous work, the diagnostic process was approximated as not only unobserved but also unchanging between observation epochs. Approximating $\{Z_t, k\Delta < t \le (k+1)\Delta\}$ with the single value $Z_{k\Delta}$ allowed a deterministic evaluation of

$$R(k, Z_{k\Delta}, t) \equiv P\left(T > k\Delta + t \mid T > k\Delta, Z_{k\Delta}\right) = \exp\left(-\psi(Z_{k\Delta})\int_{k\Delta}^{k\Delta+t} h_0(s)ds\right).$$

An attempt to apply that formula and others based on the same approximation resulted in the average replacement cost of the "optimal" replacement policy decreasing with Δ, suggesting that less frequent observations would enable better replacement decisions. This counter-intuitive result motivated the more detailed analysis in the next three sections of this paper.

2.3 Optimal Replacement Policy for Periodic Monitoring

The form of an optimal replacement policy, which minimizes the long-run expected average replacement cost per unit time for systems in the PHM with fixed Δ, was derived by Makis and Jardine (1992) while the computation of the optimal policy parameters was simplified by the discrete-time approximation of *Z*. To compare costs under different values of Δ while considering the fact that the *Z* process may change state at any time, we find the parameters of an optimal replacement policy and its cost without the discrete-time approximation, given that the form of the replacement policy follows variant 2 of the policy in Makis and Jardine (1992); that is, the system may be replaced preventively either at an observation epoch or immediately if it fails between observation epochs.

As in Makis and Jardine (1992), let decision 0 represent immediate replacement upon observation of the system state, and decision $+\infty$ correspond to non-replacement (i.e., wait and see). They showed that an optimal replacement policy δ for variant 2 exists and has the following form

$$\delta(k, z) = \begin{cases} +\infty \text{ if } K\left[1 - R(k, z, \Delta)\right] < g\int_0^{\Delta} R(k, z, t)dt \\ 0 \text{ otherwise,} \end{cases}$$

where g is the optimal average replacement cost per unit time, k is the number of monitoring intervals since the last replacement and $z = Z_{k\Delta}$ is the condition of the system at age $k\Delta$. This conclusion still holds upon substitution of $R(k, z, t)$ by $\bar{R}(k, z, t)$ in the analysis.

The optimal replacement policy δ is monotonic in the system age and state. It specifies that if the value of g were known and no failure would occur, then the optimal replacement time for a specific condition *z* would be $k_z\Delta$, where k_z is the minimum integer that satisfies the inequality:

$$K\left[1-\bar{R}(k_z,z,\Delta)\right] < g\int_0^{\Delta}\bar{R}(k_z,z,t)dt\,. \tag{2.2}$$

On the other hand, if the system fails before $k_z\Delta$, then it is replaced immediately upon failure.

According to Makis and Jardine (1992), the following algorithm may be employed to find g. Define

$$\phi(d)=\left[C+KP\left(T_d\ge T\right)\right]\Big/E\left[\min\left\{T,T_d\right\}\right] \tag{2.3}$$

where T_d is the planned replacement time associated with the expected average cost d. Here, under a given replacement policy δ_d, $P\left(T_d\ge T\right)$ is the probability of failure replacement and $E\left[\min\left\{T,T_d\right\}\right]$ is the mean replacement time considering failure. Thus, according to the theory of renewal reward processes (Ross, 2003), $\phi(d)$ is the long-run expected average cost per unit time for policy δ_d.

The algorithm is based on a fixed point result that for any $d_0\ge 0$, if $d_m=\phi(d_{m-1})$, $m=1,2,\ldots$, then $\lim_{m\to+\infty} d_m = g$. It may be described as the following procedure:

Algorithm 2-1

1 Initialize the iteration counter $m=0$, choose an arbitrary replacement policy, and set d_0 equal to the cost of the chosen policy.

2 For d_m, use (2.2) to find the planned replacement time $k_i\Delta$ associated with current system condition *i*, i.e.,

$$k_i=\min\left\{k\ge 0: K\left[1-E\left(\bar{R}(k,i,\Delta)\right)\right]\ge d_m\int_0^{\Delta}E\left(\bar{R}(k,i,t)\right)dt\right\},\ i\in S\,. \tag{2.4}$$

3 Use the replacement policy obtained in step 2 and equation (2.3) with $d=d_m$ to calculate $d_{m+1}=\phi(d_m)$.

4 If $d_{m+1} = d_m$, stop with $g = d_m$; otherwise, set $m \leftarrow m+1$ and go to step 2.

Actually, Algorithm 2-1 is an example of the policy iteration algorithm as discussed by Tijms (1986), who proved that the sequence of d values obtained from a policy improvement algorithm is monotonically decreasing and therefore the algorithm will converge in a finite number of iterations.

For step 1, a good initial choice is $d_0 = (C+K)/E(T)$, which is the long-run average cost of the policy that replaces only at failure. The crucial steps of this iteration procedure are steps 2 and 3; that is, to use equation (2.4) to identify current parameters of the replacement policy and then use equation (2.3) to update $d_{m+1} = \phi(d_m)$. The difficulties arise from the calculation of $\bar{R}(k,i,t)$ and the computation of $E(\min\{T,T_d\})$ and $P(T_d \geq T)$ under a given replacement police δ_d. In the next two sections, we will derive formulas for computing $\bar{R}(k,i,t)$, $E(\min\{T,T_d\})$ and $P(T_d \geq T)$ by conditioning.

2.4 Analysis of the Expected Conditional Reliability Function

2.4.1 Definitions

Here we introduce some new definitions to facilitate the presentation of our method. Based on the assumption and notation in section 2.2, the sojourn time X_k is exponentially distributed with rate v_k and the X_k's are mutually independent. For convenience, define $X_{n-1} \equiv +\infty$ associated with the absorbing state n-1.

For $j \geq 0$ and $i \in S$, given that the age of the system is $j\Delta$ and $Z_{j\Delta} = i$, define

$$S_{ir} = \sum_{k=i}^{r} X_k, \; r \in S \text{ and } r \geq i.$$

Then $j\Delta + S_{ir}$ is the time point that the Z process makes a transition from state i to state $r+1$. Therefore, if $t \in [S_{i,r-1}, S_{ir})$, then $Z_{j\Delta+t} = r$. For convenience, we also define $S_{i,i-1} \equiv 0$, $S_{i,n-1} \equiv +\infty$.

Define $T_R = T - j\Delta$, which is the residual time to failure if no preventive replacement is made. (Note that, for simplicity, dependence on *j* is suppressed in the notation for S_{ir} and T_R).

Then from the expected conditional reliability function (2.1), it follows that:

$$\bar{R}(j, Z_{j\Delta}, t) = E\left[P(T_R > t \mid j\Delta, Z_{j\Delta})\right] = E\left[\exp\left(-\int_{j\Delta}^{j\Delta+t} h_0(s)\psi(Z_s)ds\right) \mid Z_{j\Delta}\right]. \quad (2.5)$$

Next, we evaluate $\bar{R}(j,i,t)$ by conditioning on $S_{ii}, S_{i,i+1}, ..., S_{i,n-2}$. To better illustrate this procedure, first we examine a simple situation where the *Z* process has only three states $\{0,1,2\}$. Then we generalize the formulation of the three-state *Z* process to that of an *n*-state pure birth process.

2.4.2 Derivation of $\bar{R}(j,i,t)$ for Three-State *Z* process

As mentioned by Makis *et al*. (2003), a diagnostic process with three working states often is practical; e.g., one can view state 0 as a new system, state 1 as having some deterioration and state 2 as a warning state. Thus, it is helpful to detail the analysis for a three-state *Z* process for both illustrative and practical purposes.

Here, we analyze $\bar{R}(j,0,t)$ only**.** The formulas for $\bar{R}(j,i,t), i = 1,2$, may be deduced similarly and we relegate them to Appendix 2.A.

For a three-state *Z* process, we can evaluate $\bar{R}(j,0,t)$ by conditioning on S_{00} and S_{01}. Using the law of total expectation, we have

$$\begin{aligned}\bar{R}(j,0,t) &= E\left[P(T > j\Delta + t \mid T > j\Delta, Z_{j\Delta} = 0)\right] \\ &= E\left[E\left[P(T > j\Delta + t \mid T > j\Delta, Z_{j\Delta} = 0, S_{00}, S_{01})\right]\right].\end{aligned}$$

(2.6)

Given $Z_{j\Delta} = 0$ and for a given $t > 0$, the feasible region of the two-dimensional (S_{00}, S_{01}) space could be divided into 3 sub-regions (cases), as shown in Figure 2-1; that is, Case 0: $S_{00} \geq t$, Case 1: $S_{00} < t \leq S_{01}$ and Case 2: $S_{01} < t$.

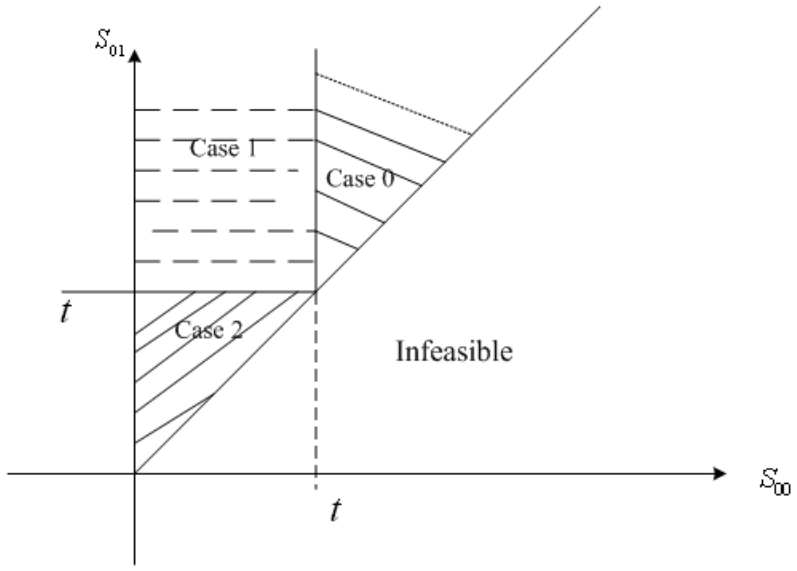

Figure 2-1 S_{ir} Space partition

Let s_{ir} represent a value (realization) of S_{ir}. Then define conditional cumulative distribution functions (CDF's) of T_R corresponding to the three cases above when $Z_{j\Delta} = 0$.

For $t \le s_{00}$,

$$F_0^0(j,t) = P(T_R \le t \mid S_{00} = s_{00}, S_{01} = s_{01}, j\Delta, Z_{j\Delta} = 0) = 1 - \exp\left(-\psi(0)\int_{j\Delta}^{j\Delta+t} h_0(u)\,du\right). \quad (2.7)$$

For $s_{00} < t \le s_{01}$,

$$\begin{aligned} F_1^0(j,t,s_{00}) &= P(T_R \le t \mid S_{00} = s_{00}, S_{01} = s_{01}, j\Delta, Z_{j\Delta} = 0) \\ &= 1 - \exp\left(-\psi(0)\int_{j\Delta}^{j\Delta+s_{00}} h_0(u)\,du - \psi(1)\int_{j\Delta+s_{00}}^{j\Delta+t} h_0(u)\,du\right). \end{aligned} \quad (2.8)$$

And for $t > s_{01}$,

$$\begin{aligned} F_2^0(j,t,s_{00},s_{01}) &= P(T_R \le t \mid S_{00} = s_{00}, S_{01} = s_{01}, j\Delta, Z_{j\Delta} = 0) \\ &= 1 - \exp\left(-\psi(0)\int_{j\Delta}^{j\Delta+s_{00}} h_0(u)\,du - \psi(1)\int_{j\Delta+s_{00}}^{j\Delta+s_{01}} h_0(u)\,du - \psi(2)\int_{j\Delta+s_{01}}^{j\Delta+t} h_0(u)\,du\right). \end{aligned} \quad (2.9)$$

We know that X_0 and X_1 are exponentially distributed and they are independent of each other. In addition, the event $S_{00} = s_{00}, S_{01} = s_{01}$ is equivalent to the event $X_0 = s_{00}, X_1 = s_{01} - s_{00}$. Hence, the joint density function of S_{00}, S_{01} is:

$$f(s_{00}, s_{01}) = \nu_0 e^{-\nu_0 s_{00}} \nu_1 e^{-\nu_1 (s_{01} - s_{00})} \quad (2.10)$$

Therefore, using equation (2.6) and setting the relevant integral domains according to the three sub-regions, we get

$$\begin{aligned}
&\bar{R}(j,0,t) = \int_t^\infty v_0 e^{-v_0 s_{00}} \left[1 - F_0^0(j,t)\right] ds_{00} + \int_t^\infty \int_0^t f(s_{00}, s_{01}) \left[1 - F_1^0(j,t,s_{00})\right] ds_{00} ds_{01} \\
&+ \int_0^t \int_0^{s_{01}} f(s_{00}, s_{01}) \left[1 - F_2^0(j,t,s_{00},s_{01})\right] ds_{00} ds_{01} \\
&= e^{-v_0 t} \left[1 - F_0^0(j,t)\right] + \int_0^t v_0 e^{-v_0 s_{00}} e^{-v_1 (t - s_{00})} \left[1 - F_1^0(j,t,s_{00})\right] ds_{00} \\
&+ \int_0^t \int_0^{s_{01}} v_0 e^{-v_0 s_{00}} v_1 e^{-v_1 (s_{01} - s_{00})} \left[1 - F_2^0(j,t,s_{00},s_{01})\right] ds_{00} ds_{01}.
\end{aligned} \tag{2.11}$$

2.4.3 Derivation of $\bar{R}(j,i,t)$ for an *n*-State *Z* process

In the situation where the Z process has n states $\{0,1,...,n-1\}$, the formulas for $\bar{R}(j,i,t)$ may be derived in the same manner as in the three-state situation. Thus, in the following, we will present the formulas for $\bar{R}(j,i,t), i = 0,1,...,n-1$, directly.

Let s_{ir} represent a value (realization) of S_{ir}. Define conditional CDF's of T_R when $Z_{j\Delta} = i$. For $s_{i,i+m-1} < t \le s_{i,i+m}$,

$$\begin{aligned}
&F_m^i(j,t,s_{ii},...,s_{i,i+m-1}) = P(T_R \le t \mid S_{ii} = s_{ii}, \cdots, S_{i,i+m} = s_{i,i+m},\, j\Delta, Z_{j\Delta} = i) \\
&= 1 - \exp\left(- \sum_{k=i}^{i+m-1} \psi(k) \int_{j\Delta + s_{i,k-1}}^{j\Delta + s_{i,k}} h_0(u)du - \psi(i+m) \int_{j\Delta + s_{i,i+m-1}}^{j\Delta + t} h_0(u)du \right), m = 0,1,...,n-i-1.
\end{aligned} \tag{2.12}$$

The joint density function of $S_{ii}, S_{i,i+1}, ..., S_{i,i+m}$ is

$$f(s_{ii}, s_{i,i+1}, ..., s_{i,i+m}) = v_i e^{-v_i s_{ii}} v_{i+1} e^{-v_{i+1}(s_{i,i+1} - s_{ii})} ... v_{i+m} e^{-v_{i+m}(s_{i,i+m} - s_{i,i+m-1})} \tag{2.13}$$

for all $m = 0,1,...,n-i-2$.

Thus,

$$\begin{aligned}
&\bar{R}(j,i,t) \\
&= \sum_{m=0}^{n-i-2} \int_t^\infty \int_0^t \int_0^{s_{i,i+m-1}} ... \int_0^{s_{i,i+1}} f(s_{ii},...,s_{i,i+m}) \left(1 - F_m^i(j,t,s_{ii},...,s_{i,i+m-1})\right) ds_{ii} ... ds_{i,i+m-2} ds_{i,i+m-1} ds_{i,i+m} \\
&+ \int_0^t \int_0^{s_{i,n-2}} ... \int_0^{s_{i,i+1}} f(s_{ii},...,s_{i,n-2}) \left(1 - F_{n-i-1}^i(j,t,s_{ii},...,s_{i,n-2})\right) ds_{ii} ... ds_{i,n-3} ds_{i,n-2}
\end{aligned} \tag{2.14}$$

for all $i \in S$, where $f(s_{ii},...,s_{i,i+m})$ is given by (2.13).

2.5 Recursive Formulas for Mean Replacement Time and Failure Probability

2.5.1 Derivation of $E(\min\{T,T_d\})$ for an *n*-State *Z* process

Like $\bar{R}(j,i,t)$, the mean replacement time $E(\min\{T,T_d\})$ and failure probability $P(T_d \geq T)$ may be computed by conditioning on the variables S_{ir}. What's more, they may be calculated efficiently using recursion. Next, we derive a recursive computational procedure for $E(\min\{T,T_d\})$. The failure probability $P(T_d \geq T)$ may be treated similarly and its derivation will be presented directly in Section 2.5.2.

For a given value $d > 0$, the replacement policy δ_d may be found using (2.4). Then $k_i\Delta$ is the planned replacement time associated with the current observed system condition, i.

Let random variable

$$T(j,i) = \min\{T,T_d\} - j\Delta$$

be the residual time to replacement given that the age of the system is $j\Delta$, $Z_{j\Delta} = i$ and the replacement policy is δ_d. Define

$$W(j,i) = E\left[T(j,i)\right],$$

so that $W(0,0) = E\left(\min\{T,T_d\}\right)$. From the definitions above, it follows that

$$W(j,i) = 0 \text{, for } j \geq k_i,$$

and for $j < k_i$, we will evaluate $W(j,i)$ by conditioning on $S_{ii}, S_{i,i+1}, ..., S_{i,n-2}$. It is natural to assume that $j < k_i$ for the remainder of this section.

Again, using the law of total expectation,

$$W(j,i) = E\left[T(j,i)\right] = E\left[E\left[\left(T(j,i) \mid S_{ii}, S_{i,i+1}, ..., S_{i,n-2}\right)\right]\right].$$

According to the state of the *Z* process at time point $(j+1)\Delta$, there are $(n-i)$ cases:

Case *m*: $Z_{(j+1)\Delta} = i + m$, that is $S_{i,i+m-1} < \Delta \leq S_{i,i+m}$

$$T(j,i)=\begin{cases}T_R \text{ if } T_R \le \Delta \\ \Delta + W\left(j+1, i+m\right) \text{ if } T_R > \Delta\end{cases}$$

where $m = 0,1,...,n-i-1$.

Then for $s_{i,i+m-1} < \Delta \le s_{i,i+m}$, define:

$$\begin{aligned} &W_m^i(j, s_{ii}, s_{i,i+1},..., s_{i,i+m-1}) = E\left(T(j,i) \mid S_{ii} = s_{ii}, S_{i,i+1} = s_{i,i+1},..., S_{i,i+m} = s_{i,i+m}\right) \\ &= \sum_{k=i}^{i+m-1} \int_{s_{i,k-1}}^{s_{ik}} t dF_k^i(j,t,s_{ii},..,s_{i,k-1}) + \int_{s_{i,i+m-1}}^{\Delta} t dF_m^i(j,t,s_{ii},..,s_{i,i+m-1}) \\ &+(\Delta + W(j+1,i+m))(1-F_m^i(j,t,s_{ii},..,s_{i,i+m-1})),\ m=0,1,...,n-i-1. \end{aligned} \tag{2.15}$$

Note from (2.15) that the conditional value of $T(j,i)$ is obtained in terms of $W\left(j+1,i+m\right), m = 0,1,...,n-i-1$. Thus this is a recursive expression.

To sum up above, we have

$$\begin{aligned} W(j,i) = &\sum_{m=0}^{n-i-2} \int_{\Delta}^{\infty} \int_{0}^{\Delta} \int_{0}^{s_{i,i+m-1}} .. \int_{0}^{s_{i,i+1}} f(s_{ii},..., s_{i,i+m}) W_m^i ds_{ii} ... ds_{i,i+m-2} ds_{i,i+m-1} ds_{i,i+m} \\ &+ \int_{0}^{\Delta} \int_{0}^{s_{i,n-2}} ... \int_{0}^{s_{i,i+1}} f(s_{ii},..., s_{i,n-2}) W_{n-1}^i ds_{ii} ... ds_{i,n-3} ds_{i,n-2} \end{aligned} \tag{2.16}$$

where the density function $f(s_{ii},..., s_{i,i+m})$ is from (2.13) and the arguments of $W_m^i(j, s_{ii}, s_{i,i+1},..., s_{i,i+m-1})$ as shown in (2.15) have been dropped for succinctness.

2.5.2 Derivation of $P(T_d \ge T)$ for an *n*-State *Z* process

Define $Q(j,i) = P(T_d \ge T \mid (j,i))$. Then $Q(0,0) = P(T_d \ge T)$ and $Q(j,i) = 0$, for $j \ge k_i$. For $j < k_i$ and $s_{i,i+m-1} < \Delta \le s_{i,i+m}$, define

$$\begin{aligned} &Q_m^i(j, s_{ii}, s_{i,i+1},..., s_{i,i+m-1}) \\ &= E\left[P\left(T \le T_d \mid S_{ii} = s_{ii}, S_{i,i+1} = s_{i,i+1},..., S_{i,i+m} = s_{i,i+m}\right)\right] \\ &= F_m^i(j,\Delta,s_{ii},..,s_{i,i+m-1}) \\ &+Q(j+1,i+m)\left(1-F_m^i(j,t,s_{ii},..,s_{i,i+m-1})\right),\ m=0,1,...,n-i-1. \end{aligned} \tag{2.17}$$

Then we have

$$Q(j,i)=\sum_{m=0}^{n-i-2}\int_{\Delta}^{\infty}\int_{0}^{\Delta}\int_{0}^{s_{i,i+m-1}}..\int_{0}^{s_{i,i+1}} f(s_{ii},...,s_{i,i+m})Q_m^i ds_{ii}...ds_{i,i+m-2}ds_{i,i+m-1}ds_{i,i+m}$$
$$+\int_{0}^{\Delta}\int_{0}^{s_{i,n-2}}...\int_{0}^{s_{i,i+1}} f(s_{ii},...,s_{i,n-2})Q_{n-i-1}^i ds_{ii}...ds_{i,n-3}ds_{i,n-2} \quad (2.18)$$

for all $i \in S$ where $f(s_{ii},...,s_{i,i+m})$ is from (2.13) and Q_m^i is from (2.17) with arguments suppressed.

2.6 Optimal Age-Based Replacement

To investigate the value of condition monitoring, we also studied the optimal age-based replacement policy as a baseline for comparison.

Without any condition monitoring, preventive replacement would be based only on the age of the system. If $F(t)$ is the distribution function of the failure time and the system is replaced whenever it fails or reaches age τ, then one can find the average replacement rate,

$$\lambda_r(\tau)=\left[\int_0^{\tau} sf(s)ds+\tau\left(1-F(\tau)\right)\right]^{-1}=\left[\int_0^{\tau}\left(1-F(s)\right)ds\right]^{-1},$$

and the corresponding failure rate,

$$\lambda_d(\tau)=F(\tau)\lambda_r(\tau)$$

(see (Ross, 2003), p.461). The optimal replacement age, τ^*, is found by minimizing the total average cost per unit time, which is given by:

$$w(\tau)=C\lambda_r(\tau)+K\lambda_d(\tau)=\left(C+KF(\tau)\right)\left[\int_0^{\tau}\left(1-F(s)\right)ds\right]^{-1}. \quad (2.19)$$

In the notation of this paper, we have

$$F(t)=1-\bar{R}(0,0,t)$$

where $\bar{R}(0,0,t)$ is obtained from equation (2.14).

2.7 Numerical Illustration

To illustrate our model and its use in assessing the value of monitoring information, we consider the following numerical example. Assume that the baseline distribution is Weibull with hazard rate

$$h_0(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}},$$

where $\alpha = 1, \beta = 2$, and let $\psi(Z_t) = \exp(2Z_t)$, $C = 5$ and $K = 25$. Assume the stochastic process Z has three states $\{0,1,2\}$ with transition rates $v_0 = v_1 = -\ln(0.4), v_2 = 0$. Since the forms of $h_0(t)$ and $\psi(Z_t)$ are predefined, the PHM here is parametric rather than semi-parametric as described in Cox *et al*. (1984).

2.7.1 Replacement Policy under Periodic Monitoring

With $\Delta = 1$ in Algorithm I, we initialize $d_0 = (C + K) / E(T) = 46.8823$, which is the cost of the policy that replaces only at failure. Then we illustrate how the first iteration for finding g proceeds below. Other iterations are similar.

Iteration 1: $d_0 = 46.8823$. For $Z_t = i = 0$, we get $k_0 = 1$ from (2.2) and (2.14). Thus $W(1,0) = 0$, $Q(1,0) = 0$. Similarly, for $i = 1$ and $i = 2$, we get $k_1 = 1$ and $k_2 = 1$. Thus $W(1,1) = 0$, $Q(1,1) = 0$, $W(1,2) = 0$, $Q(1,2) = 0$. Based on these value, we obtain $W(1,0) = 0.5943$ from (2.16) and $Q(1,0) = 0.8410$ from (2.18).

The complete results are shown in Table 2-1. The policy iteration algorithm converges after a single iteration to the optimal average cost $g = 43.7905$. The algorithm was implemented in *Mathematica*® for precise and efficient numerical evaluation of multiple integrals.

Table 2-1 An Illustration of the Computation Procedure (three states)

d	k_0	k_1	k_2	$W(0,0)$	$Q(0,0)$	$\phi(d)$
46.8823	1	1	1	0.5943	0.8410	43.7905
43.7905	1	1	1	0.5943	0.8410	43.7905

To study the effect of the interval between observations, we varied Δ from 0.001 (to approach the case with continuous monitoring) to 10 (to approximate the situation without monitoring). Table 2-2 shows the optimal policies and replacement costs for various values

of Δ with three-state Z process. Notably, if no preventive replacement is done, the mean time to failure of the system may be obtained from

$$E(T) = \int_0^\infty \bar{R}(0,0,t)dt = 0.6399, \tag{2.20}$$

which agrees with the value of $W(0,0)$ when $\Delta = 10$. Table 2-2 indicates that as the inspection interval Δ decreases, the optimal replacement cost also decreases. This result is expected because with smaller Δ values we obtain more timely information about the system, and thus can respond to condition deterioration more promptly.

Table 2-2 Effect of Changing Δ on the Optimal Policy and Cost with Comparison to Age-Based Replacement

Δ	k_0	k_1	k_2	$W(0,0)$	$Q(0,0)$	g_Δ	m^*	$w(m^*\Delta)$
0.001	487	66	9	0.3690	0.1606	24.4286	285	32.4929
0.01	48	6	1	0.3664	0.1616	24.6698	29	32.4972
0.05	9	1	1	0.3553	0.1658	25.7381	6	32.5318
0.1	4	1	1	0.3329	0.1602	27.0455	3	32.5318
0.2	2	1	1	0.3444	0.2062	29.4829	2	34.0449
1	1	1	1	0.5943	0.8410	43.7905	1	43.7905
10	1	1	1	0.6399	1.0000	46.8844	1	46.8844

However, the opposite behavior occurred when we applied the discrete approximation formulas from Makis and Jardine (1992) directly to acquire the optimal policies. The results are shown in Table 2-3. To apply their discrete-time formulas, by uniformization we converted the continuous time Markov chain Z discussed above to a discrete-time Markov chain, which makes a transition every Δ units of time and has the transition probability matrix

$$P = \begin{bmatrix} 0.4^\Delta & 1-0.4^\Delta & 0 \\ 0 & 0.4^\Delta & 1-0.4^\Delta \\ 0 & 0 & 1 \end{bmatrix},$$

and we assume that all else are held equal.

Since we ignored possible transitions between inspection intervals, there is no wonder that the results in Table 2-3 are all overoptimistic, that is, for the same Δ, the optimal replacement cost in Table 2-3 is smaller than that in Table 2-2. One apparent problem of Table 2-3 is that as Δ increases from 0.001 to 0.2, the optimal replacement cost unexpectedly decreases. (We expected the optimal replacement cost to increase with Δ because less frequent observations lead to less information available, based on which it is impossible to make better decisions.) Another problem is that the average replacement time $W(0,0)$ when $\Delta = 1$ or $\Delta = 10$ is larger than the mean time to failure of the system (2.20). Despite these drawbacks, the results for $\Delta = 0.001$ indicate that the discrete-version formulas from Makis and Jardine do provide an accurate approximation for the continuous time model when Δ is sufficiently small.

Table 2-3 Optimal Policies of Various Δ According to Makis and Jardine (1992)

Δ	k_0	k_1	k_2	$W(0,0)$	$Q(0,0)$	g
0.001	488	66	9	0.3695	0.1606	24.3967
0.01	49	7	1	0.3720	0.1624	24.3503
0.05	10	1	1	0.3821	0.1692	24.1569
0.1	5	1	1	0.3907	0.1734	23.8946
0.2	2	1	1	0.3491	0.1819	23.6061
1	1	1	1	0.7468	0.6321	27.8553
10	1	1	1	0.8862	1	33.8514

2.7.2 Comparison with Age-Based Replacement

To weigh the benefit of condition information against its cost, we can compare the optimal replacement cost of the policy based on more or less frequent monitoring to that of the age-based replacement policy. We also compute the optimal age-based replacement policy, shown with its cost in the last two columns of Table 2-2. The optimal replacement age, τ^*, is found numerically by minimizing (2.19) using a heuristic search technique. To compare with the condition-based replacement policy, we constrain it to be an integer multiple, m^*, of Δ. The numerical results quantify the savings $w(m^*\Delta) - g$ that are obtained with small values of Δ by having access to more frequent observations of the product's

condition. These cost savings could justify the investment in equipment and software required to monitor the condition frequently.

The additional cost of a failure replacement, K, is usually difficult to estimate. But it could be very high for critical systems, often several times bigger than the regular replacement cost. Table 2-4 shows the impact of this cost on the optimal replacement policy and average cost when $\Delta = 0.01$. As expected, for larger values of K, the cost savings $w(m^*\Delta) - g$ provided by condition monitoring is more substantial, which implies the great importance of the condition information in critical systems.

Table 2-4 Effect of Increasing K on the Optimal Policy and Cost when $\Delta = 0.01$ with Comparison to Age-Based Replacement

K	k_0	k_1	k_2	$W(0,0)$	$Q(0,0)$	g	m^*	$w(m^*\Delta)$
$K = 5C = 25$	91	12	2	0.5150	0.3637	27.3659	29	32.4972
$K = 10C = 50$	33	4	1	0.2773	0.0879	33.8817	20	43.6787
$K = 20C = 100$	23	3	1	0.2052	0.0465	47.0403	15	58.4512

2.7.3 Optimal Monitoring Scheme

We compare age-based, periodic monitoring and continuous monitoring based on total average cost per unit time. Without monitoring, the optimal value of G_1 is obtained in section 2.6 by minimizing (2.19). We denote it as $G_1^* = G_1(\tau^*)$. The cost of the periodic monitoring scheme, G_2, is a function of the inspection interval, Δ. Its optimal value, denoted as $G_2^* = G_2(\Delta^*)$, is obtained numerically by searching the Δ space. The continuous monitoring cost, G_3, achieves its optimal value, G_3^*, when the system is under the optimal replacement policy of continuous monitoring, which we approximate by letting Δ approach 0. If $G_3^* = \min\{G_1^*, G_2^*, G_3^*\}$, then a one-time investment in continuous monitoring is worthwhile. Similarly, a smaller value of G_1^* than both G_2^* and G_3^* means that it is not worthwhile to devote any effort to collecting information on the system condition. This case can happen if

the covariates we study have an insignificant influence on the system hazard rate or the cost ratio $(C+K)/C$ is small. The optimal monitoring scheme is therefore determined by comparison among the values of G_1^*, G_2^*, G_3^*.

In our numerical example of Table 2-2, we have $G_1^* = 32.4929$ and $G_3^* = 24.4286 + \Gamma'$ (approximating g_0 as $\hat{g}_0 = g_{0.001}$). For simplicity, we restrict the value of Δ to a finite set $\Lambda = \{0.01, 0.05, 0.1, 0.2, 1, 10\}$. Then

$$G_2^* = \min_{\Delta \in \Lambda} \left(g_\Delta + \frac{\gamma}{\Delta} \right).$$

Figure 2-2 displays a plot $G_2(\Delta) - G_1^*$ to compare between G_2^* and G_1^*. The contour of G_2^* is highlighted with bold black. It is clear that if γ is smaller than approximately 0.6 (exact value is 0.6020), we can choose a proper Δ to make the periodic monitoring scheme better than no monitoring.

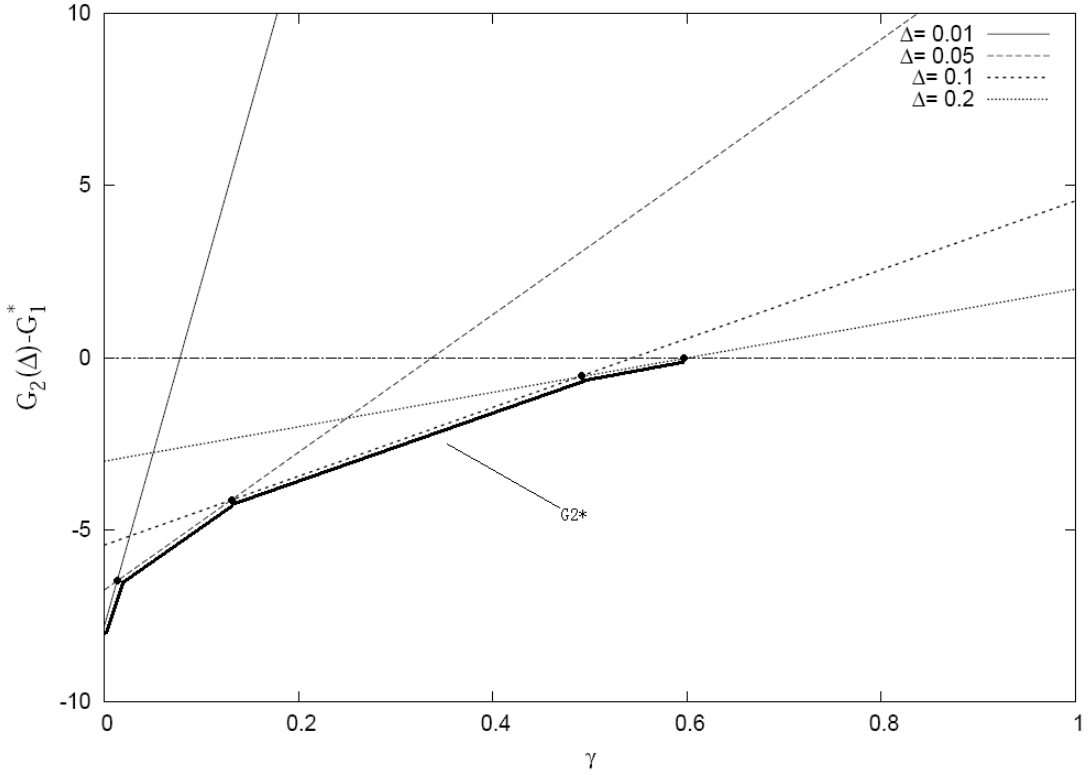

Figure 2-2 Comparison between G_1^* and G_2^*

We would like to know under what conditions the continuous monitoring scheme would be the best option. Clearly, $\Gamma' \le w\left(m^*\left(0.001\right)\right) - g_{0.001} = 8.0643$ is necessary for $G_3^* \le G_1^*$. Besides that, when $\gamma \le 0.6020$, for $G_3^* \le G_2^*$ we must have:

- if $0.4875 < \gamma \le 0.6020$, then $\Gamma' \le \gamma/0.2 + g_{0.2} - \hat{g}_0 = 5\gamma + 5.0543$;
- if $0.1307 < \gamma \le 0.4875$, then $\Gamma' \le \gamma/0.1 + g_{0.1} - \hat{g}_0 = 10\gamma + 2.6169$;
- if $0.0134 < \gamma \le 0.1307$, then $\Gamma' \le \gamma/0.05 + g_{0.05} - \hat{g}_0 = 20\gamma + 1.3095$
- if $\gamma \le 0.0134$, then $\Gamma' \le \gamma/0.01 + g_{0.01} - \hat{g}_0 = 100\gamma + 0.2412$.

This analysis indicates that when it comes to choosing a proper monitoring scheme for a specific system, it is important to weigh the benefit of monitoring against its cost carefully. Although condition-based maintenance often leads to a lower cost than age-based maintenance, this is not always the case. In our numerical example, the combinations of monitoring costs γ and Γ' under which the different monitoring schemes are optimal are shown in Figure 3. Note that the boundary between continuous and periodic monitoring could be described as the critical $r\Gamma$ being a concave piecewise-linear function of γ. This occurred when we restricted the value of Δ to a finite set; we conjecture that if the value of Δ is allowed to vary continuously, the critical $r\Gamma$ would be a smooth increasing concave function of γ. One implication of this concave shape is as follows. Suppose that current costs lie in the region where periodic monitoring is optimal; i.e., the initial cost, Γ, to set up continuous monitoring is prohibitively expensive relative to the periodic monitoring cost, γ. If γ increases, for example due to growth in labor costs, then the drop in Γ required to make continuous monitoring worthwhile becomes disproportionately smaller.

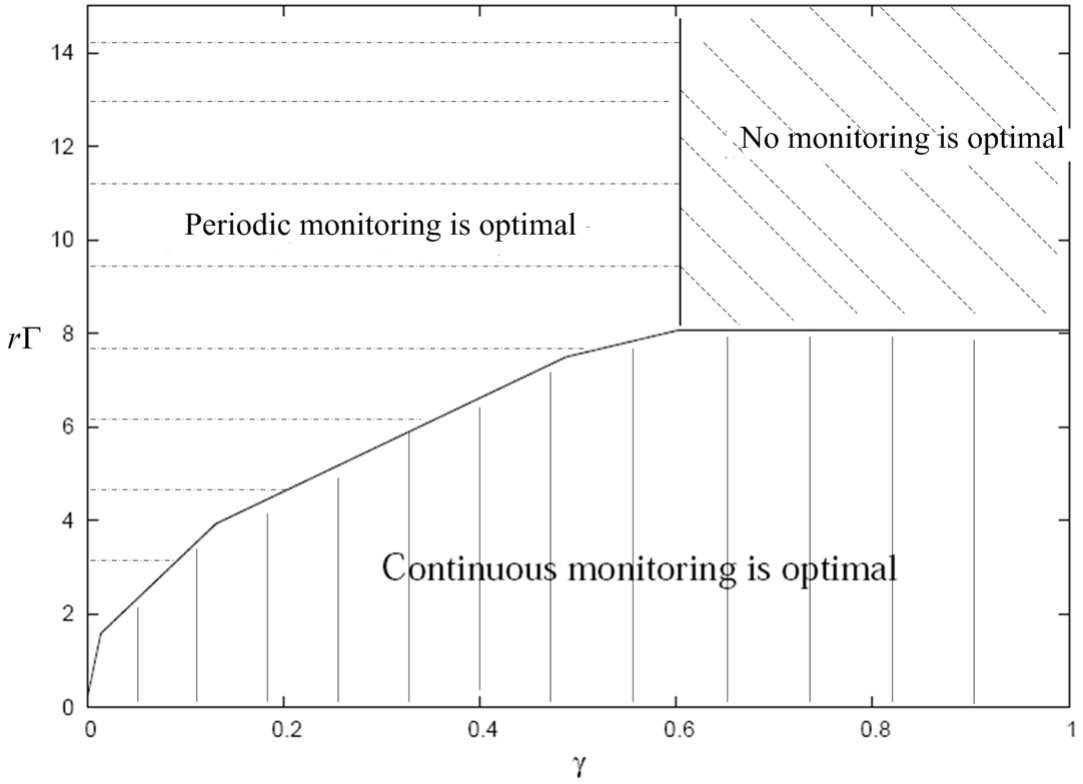

Figure 2-3 Optimal cost regions for different monitoring schemes

2.8 Conclusion

In this paper, we investigated a condition-based replacement problem under various monitoring schemes for a deteriorating system with concomitant conditions described by a continuous time Markov chain. The proportional hazards model was applied to describe the failure time of this system. For such a model, although the form of the optimal replacement policy under periodic monitoring was given by Makis and Jardine (1992), computing the optimal policy parameters for a system with a continuous time diagnostic process is delicate. First, a recursive procedure was developed to obtain the optimal average cost and the parameters of the optimal policy for system with an n-state pure birth process. Then a numerical example with n=3 illustrated the computational procedure as well as the evaluation of condition information with more or less frequent monitoring. At last by taking the monitoring cost into consideration, we obtained the relationships between the cost γ of each inspection under periodic monitoring and the upfront cost Γ of continuous monitoring, under

which the continuous, periodic or no monitoring scheme minimizes the total average cost per unit time. Specifically, in the numerical example, no monitoring (i.e., age-based replacement) is optimal if both γ and Γ exceed certain values; and, for a fixed interest rate, the critical Γ on the boundary between continuous and periodic monitoring optimality is a concave increasing function of γ.

Extensions of this research could include generalizing the one-dimensional covariate vector to multi-dimensional. Then the *Z* process would be a general Markov chain rather than a pure birth process. It could evolve along multiple paths, which would make the calculation of policy parameters by conditioning extremely intricate. In addition, the Markovian assumption of the diagnostic process could be relaxed to a semi-Markovian process, which allows arbitrary sojourn time distributions. Also in this paper, we assumed that the condition of the product is assessed perfectly, but in real situations it is only partially observed. The value of condition monitoring would be estimated more accurately by considering the element of uncertainty added by partial observations. Although Ghasemi *et al*. (2007) solved the partial observation problem on Makis and Jardine's model using dynamic programming, the approximation of the Z process as constant within inspection intervals was left intact. Further extensions could generalize the underlying failure model. Using a different model to relate the concomitant information to system failure time distribution, such as a scale-accelerated failure time (SAFT) model (Meeker and Escobar, 1998), could be of great practical value. In this case, both the optimal policy and its calculation must be reconsidered.

Appendix 2.A Formulas for $\bar{R}(j,i,t)$ with $i=1,2$ for Three-State *Z* Process

2.A.1 Formulas for $\bar{R}(j,1,t)$

Define conditional CDF's of T_R when $Z_{j\Delta}=1$. For $t \le s_{11}$, we have

$$F_0^1(j,t) = P(T_R \le t \mid S_{11}=s_{11},\, j\Delta, Z_j=1) = 1-\exp\left(-\psi(1)\int_{j\Delta}^{j\Delta+t} h_0(u)\,du\right),$$

and for $t > s_{11}$, we have

$$F_1^1(j,t,s_{11}) = P(T_R \le t \mid S_{11} = s_{11},\, j\Delta, Z_j = 1)$$
$$= 1 - \exp\left(-\psi(1)\int_{j\Delta}^{j\Delta+s_{11}} h_0(u)\,du - \psi(2)\int_{j\Delta+s_{11}}^{j\Delta+t} h_0(u)\,du\right).$$

Then we have

$$\bar{R}(j,1,t) = e^{-\nu_1 t}(1 - F_0^1(j,t)) + \int_0^t \nu_1 e^{-\nu_1 s_{11}}(1 - F_1^1(j,t,s_{11}))ds_{11}\,.$$

2.A.2 Formulas for $\bar{R}(j,2,t)$

Define conditional CDF's of T_R when $Z_{j\Delta} = 2$,

$$F_0^2(j,t) = P(T_R \le t \mid j\Delta, Z_j = 2) = 1 - \exp\left(-\psi(2)\int_{j\Delta}^{j\Delta+t} h_0(u)\,du\right).$$

Then we have

$$\bar{R}(j,2,t) = 1 - F_0^2(j,t)\,.$$

Acknowledgements

This work was supported by the National Science Foundation under grant CNS-0540293.

References

Aven, T. and Bergman, B. (1986). Optimal replacement times -- a general set-up. *Journal of Applied Probability*, 23, 432-442.

Banjevic, D., Jardine, A. K. S., Makis, V., and Ennis, M. (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR, 39,* 32-50.

Bloch-Mercier, S. (2002). A preventive maintenance policy with sequential checking procedure for a Markov deteriorating system. *European Journal of Operational Research, 142,* 548-576.

Chen, C. T., Chen, Y. W., and Yuan, J. (2003). On a dynamic preventive maintenance policy for a system under inspection. *Reliability Engineering & System Safety, 80,* 41-47.

Chiang, J. H. and Yuan, J. (2001). Optimal maintenance policy for a Markovian system under periodic inspection. *Reliability Engineering & System Safety, 71,* 165-172.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.

Dieulle, L., Berenguer, C., Grall, A., and Roussignol, M. (2003). Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of Operational Research, 150,* 451-461.

Ghasemi, S., Yacout, S., and Ouali, M. S. (2007). Optimal condition based maintenance with imperfect information and the proportional hazards model. *International Journal of Production Research, 45,* 989-1012.

Lam, C. T. and Yeh, R. H. (1994a). Comparison of Sequential and Continuous Inspection Strategies for Deteriorating Systems. *Advances in Applied Probability, 26,* 423-435.

Lam, C. T. and Yeh, R. H. (1994b). Optimal maintenance policies for deteriorating systems under various maintenance strategies. *IEEE Transactions on Reliability, 43,* 423-430.

Liao, H. T., Elsayed, A., and Chan, L. Y. (2006). Maintenance of continuously monitored degrading systems. *European Journal of Operational Research, 175,* 821-835.

Makis, V. and Jardine, A. K. S. (1992). Optimal replacement in the proportional hazards model. *INFOR, 30,* 172-183.

Makis, V. and Jiang, X. (2003). Optimal replacement under partial observations. *Mathematics of Operations Research, 28,* 382-394.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data,* Wiley, New York.

Rosenblatt, M. J. and Lee, H. L. (1986). A comparative study of continuous and periodic inspection policies in deteriorating production systems. *IIE Transactions, 18,* 2-9.

Ross, S. M. (2003). *Introduction to Probability Models*. (8th ed.), Academic Press, San Diego, CA.

Tijms, H. C. (1986). *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, New York.

Yeh, R. H. (1997). Optimal inspection and replacement policies for multi-state deteriorating systems. *European Journal of Operational Research, 96,* 248-259.

CHAPTER 3 OPTIMAL REPLACEMENT IN THE PROPORTIONAL HAZARDS MODEL WITH SEMI-MARKOVIAN COVARIATE PROCESS AND CONTINUOUS MONITORING

A paper published in *IEEE Transactions on Reliability*[2]

Xiang Wu and Sarah M. Ryan

Abstract

Motivated by the increasing use of condition monitoring technology for electrical transformers, this paper deals with the optimal replacement of a system having a hazard function that follows the proportional hazards model with a semi-Markovian covariate process, which we assume is under continuous monitoring. Although the optimality of a threshold replacement policy to minimize the long-run average cost per unit time was established previously in a more general setting, the policy evaluation step in an iterative algorithm to identify optimal threshold values poses computational challenges. To overcome them, we use conditioning to derive an explicit expression of the objective in terms of the set of state-dependent threshold ages for replacement. The iterative algorithm is customized for our model to find the optimal threshold ages. A three-state example illustrates the computational procedure, as well as the effects of different sojourn time distributions of the covariate process on the optimal policy and cost. Numerical examples and sensitivity analysis provide some insights into the suitability of a Markov approximation, and the sources of variability in the cost. The optimization method developed here is much more

[2] Appeared in *IEEE Transactions on Reliability*, 2011, 60, 580-589

efficient than the approach that approximates continuous monitoring as periodic, and then optimizes the periodic monitoring parameters.

Index terms—Optimal replacement, proportional hazards model, semi-Markov process, threshold replacement policy, sensitivity analysis

ACRONYM

CBM	Condition-based maintenance
DGA	Dissolved gas analysis
PHM	Proportional hazards model
STD	Sojourn time distributions
CV	Coefficient of variation

NOTATION

t	The age of the current system.
$Z = \{Z_t, t \geq 0\}$	A right continuous semi-Markov process with a finite state space $\{0,1,...,n-1\}$ and $Z_0 = 0$ that reflects the health condition of the system at age *t*.
$h_0(t)$	The baseline hazard rate, which depends only on the age of the system.
$\psi(Z_t)$	The link function in PH model that depends on the state of the covariate process Z.
X_k	The sojourn time of the *Z* process in state k, $k = 0,...,n-2$.
$f_{X_k}(x_k)$	The pdf of X_k, $k = 0,...,n-2$.
S_k	The age at which the covariate state changes from k to $k+1$, $k = 0,...,n-2$.
$g_k(s_0, s_1, ..., s_k)$	The joint pdf of $S_0, S_1, ..., S_k$, $k = 0,...,n-2$.
$G_k(s_0, s_1, ..., s_k)$	The joint Cdf of $S_0, S_1, ..., S_k$, $k = 0,...,n-2$.
T	The time to failure of the system.

T_d	A stopping time dependent on the age of the system and Z_t.
δ_{T_d}	A replacement policy that replaces at failure or at T_d, whichever occurs first.
C	The replacement cost without failure, $C > 0$.
K	The additional cost for a failure replacement, $K > 0$.

ASSUMPTIONS

1. The system must be kept in working order at all times. Replacement is instantaneous.
2. The baseline hazard rate, $h_0(t)$, is a non-decreasing function of the system age; that is, the system deteriorates with time.
3. The link function, $\psi(Z_t)$, is a non-decreasing function with $\psi(0) = 1$.
4. The practice of continuous monitoring influences neither the covariate process Z nor the system failure process.

3.1 Introduction

This article concerns a condition-based maintenance (CBM) problem for critical assets. Compared to classical preventive maintenance, CBM improves the decision-making process by exploiting available information about the system's operating conditions. Increasingly, condition monitoring technology is gaining favor as a way to diagnose the health status, and detect the impending failure of expensive assets.

This work was motivated by the need to improve the management of capital-intensive assets such as high-voltage power transformers. As explained by Wang et al. [1], "As transformers age, their internal condition degrades, which increases the risk of failure. Failures are usually triggered by severe conditions, such as lightning strikes, switching transients, short-circuits, or other incidents. When the transformer is new, it has sufficient electrical and mechanical strength to withstand unusual system conditions. As transformers age, their insulation strength can degrade to the point that they cannot withstand system events such as short-circuit faults or transient overvoltages." Unexpected failure of power

transformers results in unscheduled outages with power delivery problems, and may cause immense economic loss. For example, the replacement cost of a single phase 500 MVA transformer is around 1 million dollars, while the failure cost could run several times as high as that number [2]. To reduce the risk of unexpected failure, on-line monitoring has become common practice, and the condition information concerning transformers in the field can be returned in real time to a central location for continuous assessment [2], [3]. A real example is described in [2], where on-line dissolved gas analyzers attached to transformers collect dissolved gas analysis (DGA) data six times each day, on a regular 4-hour schedule. In view of the multi-decade life cycle of transformers, it is adequate to view this kind of practice as continuous monitoring. The high cost of unexpected transformer failure motivates our study of how to make best use of the condition information to decide when to perform preventive replacement.

CBM models of the system's lifetime differ according to their approaches of utilizing the condition information. Many researchers assume that the system failure process can be described adequately by a multi-state deteriorating model, and extensive research has been done with Markov and semi-Markov decision models [4] - [9]. Douer and Yechiali [6] studied the optimal repair and replacement problem in Markovian systems, and they introduced a generalized control limit policy which is optimal under reasonable conditions. Lam and Yeh [7] used a semi-Markov process to model a multi-state deteriorating system, and considered state-age-dependent replacement policies. They showed that optimal replacement policies have monotonic properties under reasonable assumptions on replacement cost, replacement time, and failure rate. Chen and Trivedi [8] built a semi-Markov decision model for condition-based maintenance policy optimization, and presented an approach to optimize the inspection rate and maintenance policy jointly. The issues of imperfect monitoring in state-based preventive maintenance were considered in [10], [11]. In contrast to the multi-state deteriorating models, Toscano and Lyonnet [12] proposed a dynamic failure rate model that predicts the reliability of the system in real time by taking into account the past and present operating conditions.

Another valuable and increasingly prevalent way to incorporate condition information into risk estimation is the proportional hazards (PH) model [13], which explicitly includes

both the age and the condition information in the calculation of the hazard function. It combines a baseline hazard function which accounts for the aging degradation with a link function that takes the condition information into account to improve the prediction of failure. Generally, the condition information is described by a multi-state covariate (diagnostic) process $Z = \{Z_t, t \geq 0\}$. The PH-based replacement policies have been successfully applied in a variety of industrial sectors such as pulp and water, coal plants, nuclear plant refueling, military land armored vehicles, construction industry backhoes, marine diesel engines, and turbines in a nuclear plant [14].

Several papers have been published to optimize the decision-making in the PH model setting. Makis and Jardine [15] investigated the optimal replacement policy for systems under a PH model with a Markov covariate process, and periodic monitoring; and they showed that the optimal replacement policy is of a control limit type in terms of the hazard function. Banjevic and Jardine [16] extended Makis and Jardine's model by relaxing the monotonicity assumption of the hazard function, and they developed methods for parameter estimation in the PH model as well. The same model was extended in [17] by assuming the information obtained at inspection epochs is imperfect; that is, the condition information of the system is only partially observed. Wu and Ryan [18] removed the discrete-time approximation of the continuous time covariate process in [15], which could lead to a counter-intuitive result when comparing the cost of policies with different monitoring intervals. They presented a new recursive procedure to obtain the optimal policy, and assess whether the investment of condition monitoring technology in capital-intensive physical assets is worthwhile. All of these papers assumed the covariate processes to be Markov processes, and under periodic monitoring.

In this paper, we extend the PH-based replacement models to systems with semi-Markov covariate processes under continuous monitoring. We consider parametric PH model with a baseline hazard function, and a time dependent covariate process. In the transformer application, it is reasonable to let the covariate process Z represent the condition of the insulation, which degrades over time, and may be classified into several different states, such as new, normal, warning, and dangerous. Assume the state of the insulation can be perfectly inferred from a combination of monitored variables including acoustic and electrical signals

caused by partial discharge, moisture or gases in the insulating oil, or other quantities that indicate the condition of the insulation [1]. By modeling the evolution of the insulation state, the hazard function for the transformer can be evaluated, and further, the mean time to failure and the average cost associated with any given replacement policy can be calculated. In the PH model setting, a transformer failure can occur from any insulation state with increasing risk of failure as the insulation condition degrades.

Maintenance to improve the condition of the insulation requires taking the transformer out of service for a significant period of time to replace the insulation, which is not a practical option. Besides, the maintenance cost is relatively low compared to the preventive replacement cost plus the failure cost [2], [19]. Thus, in this paper, we consider replacement of the transformer as the only maintenance option.

Examining existing PH-based replacement models exposed the gaps between the literature and practice. So far, the form of the optimal policy for systems under continuous monitoring has not been articulated, and how to estimate the risk with continuously monitored information has not been addressed. In addition, a Markovian model may not be appropriate for the covariate process. Requiring that times between transitions among the covariate states be exponentially distributed is an added approximation which limits the usage of the model. Therefore, we adopt a semi-Markov covariate process with general transition time distributions.

The contributions of this research and outline of the paper are as follows. By identifying our model as a special case of the one described in [20], we show in Section 3.3 that, if the hazard function of the system is non-decreasing, then the optimal replacement policy is of the control limit type with respect to the hazard function, and may be uniquely defined by a set of state-dependent threshold ages for replacement. To compute the optimal policy and optimal cost, we use conditioning arguments to derive explicit expressions for s-expected life and failure probability of transformers in terms of the policy parameters in Section 3.4. The iterative procedure developed by Bergman [20] is specified for our model to find the optimal threshold ages. The model and the solution procedure are illustrated by numerical examples in Section 3.5. We discuss its computational advantage over the recursive procedure [18], and we study the effect of different sojourn time distributions of the covariate process on the

optimal policy and cost. In addition, sensitivity analysis is performed on a specific instance to demonstrate how the variations in the input parameters would affect the long-run average cost.

3.2 Model Description

We assume the system deteriorates with time, and is subject to random failure. Upon failure, the system is instantaneously replaced by a new one, and the process renews. The hazard function of the system increases with the system's age, as well as with the value of covariates that reflect the health condition of the system.

For simplicity, we consider only one covariate. To account for both the age effect and the condition information in the system's hazard function, the PH model is employed to describe the failure process of the system. That is, the hazard function of the system at time t can be expressed as

$$h(t, Z_t) = h_0(t)\psi(Z_t), t \geq 0 \quad . \tag{3.21}$$

We assume that $Z = \{Z_t, t \geq 0\}$ is a continuous-time semi-Markov process which depicts the evolution of the covariate, and is under continuous monitoring. It has a finite state space $\{0,1,...,n-1\}$, where state 0 represents the covariate state corresponding to a new system, and states $1,2,...,n-1$ reflect the increasingly deteriorating condition. It follows that the conditional survivor function is given by

$$R(t;Z) = \Pr(T > t \mid Z_s, 0 \leq s \leq t) \equiv \exp\left(-\int_0^t h_0(s)\psi(Z_s)ds\right), t \geq 0. \tag{3.22}$$

From this function, we can see that a system failure can occur in any state at any time with increasing likelihood as the system ages, and the health condition degrades.

Between any two consecutive replacements, the covariate process Z changes states according to a pure birth process; i.e., whenever a transition occurs, the state of the process always increases by one, and state $n-1$ is absorbing. Replacement is instantaneous, and the covariate returns to state 0 upon replacement. The time interval between two successive transitions is a random variable with any distribution. Let X_k be the sojourn time in state k. We allow X_k to follow an arbitrary distribution with density $f_{X_k}(x_k)$, for $k \leq n-2$; the

distribution of X_{n-1} is immaterial because the covariate process exits from that state only when the system is replaced. Define $S_k = \sum_{i=0}^{k} X_i, k = 0,1,..,n-2$, which is the age when the covariate moves from state k to state $k+1$. The joint pdf, and Cdf of $S_0, S_1, ..., S_k$, for $k = 0,...,n-2$, are represented as $g_k(s_0, s_1, ..., s_k)$, and $G_k(s_0, s_1, ..., s_k)$ respectively, where $0 < s_0 < s_1 < ... < s_k$. As will be shown in Section 4, the pdf $g_k(s_0, s_1, ..., s_k)$ is fully determined by $f_{X_k}(x_k)$, $k = 0,...,n-2$, as is $G_k(s_0, s_1, ..., s_k)$.

In practice, the state of the covariate is inferred from continuously monitored variables. In the transformer application, the state of insulation is determined by a combination of acoustic and electrical signals, detection of moisture or gases in the insulating oil, dissolved gas analysis data, and so on. By carefully examining historical data, the point in time at which the covariate changes state would be known, and the forms and the parameters of $f_{X_k}(x_k)$ could be identified and estimated using standard statistical methods.

Continuous monitoring usually involves an upfront investment in hardware and software installation, and each inspection action costs nothing thereafter. Because this upfront cost does not affect the optimal policy that minimizes the long-run average cost, we do not include the cost of continuous monitoring in our objective function.

Define the replacement rule δ_{T_d} : Replace at failure or at T_d, whichever occurs first. Utilizing the classical cost structure, assume each planned replacement costs $C > 0$, and each failure replacement incurs an additional cost $K > 0$. Then, according to the theory of renewal reward processes [21], the long run average cost per unit time can be expressed as

$$\phi(T_d) = \frac{C + K \Pr(T_d \geq T)}{E\left[\min\{T, T_d\}\right]} \tag{3.23}$$

where $\Pr(T_d \geq T)$ is the probability of failure replacement, and $E\left[\min\{T, T_d\}\right]$ is the *s*-expected replacement time. The main objective of this paper is to find an optimal replacement policy that minimizes the long-run average cost per unit time for systems with semi-Markovian covariate process, and continuous inspection; and to establish procedures to obtain the parameters of the optimal policy.

3.3 The Form of the Optimal Replacement Policies

Bergman [20] investigated the optimal replacement problem under a general failure model, in which the hazard rate $h(\cdot)$ of system failure is non-decreasing, and completely determined by a general stochastic process $X(t)$, $t \geq 0$. It is assumed that $X(t)$ is also non-decreasing, and under continuous monitoring. Under the same cost structure as in Section 2, Bergman showed that the optimal replacement policy is of the control limit type, and the optimal stopping time has the form

$$T_d^* = \inf\{t \geq 0 : h\big(X(t)\big) \geq d^* / K\} \tag{3.24}$$

where $d^* = \phi(T_d^*)$ is the optimal cost. If the set in (3.24) is empty, then $T_d^* = \infty$, which means replacement only at failure.

Equation (3.24) indicates that, for a given control limit d^* / K, the optimal policy parameters can be calculated. However, d^* itself is dependent on the optimal policy. To solve this difficulty, Bergman proved the following proposition, which leads to an iterative algorithm that produces a sequence converging to an optimal cost.

Proposition 3-1. *Choose any positive* d_0, *and set iteratively*

$$T_n = \inf\{t \geq 0 : h\big(X(t)\big) \geq d_n / K\} \tag{3.25}$$

$$d_{n+1} = \phi(T_n), \quad n = 0,1,2,\ldots. \tag{3.26}$$

Then $\lim_{n\to\infty} d_n = d^*$.

A generalization made in the latter part of [20] greatly extends the application scope of this model. Therein Bergman stated that the process $X(t)$ can be generalized to be a stochastic vector process with $X = (X_1, X_2, ..., X_n)$, which represents n different measurements of deterioration. As long as each component of $X(t)$ is non-decreasing, and the state-dependent hazard rate function $h\big(X(t)\big)$ is non-decreasing in each component of $X(t)$, the above conclusions hold.

The PH model with a semi-Markovian covariate process and continuous monitoring presented in Section 2 is a special case of the general failure model defined by Bergman, where the age of the system could be regarded as one component of the stochastic process

$X(t)$, and the covariate Z_t as the other component of $X(t)$. Thus we obtain the following theorem.

Theorem 3-1. *For a system whose failure time follows the proportional hazards model* (1) *that is to be replaced at the smaller of its failure time or a replacement stopping time, the optimal stopping time satisfies*

$$T_d^* = \inf\{t \geq 0, h_0(t)\psi(Z_t) \geq d^* / K\} \tag{3.27}$$

where d^* *is the optimal cost.*

The optimal replacement policy specified by (3.27) may be explained as: *replace at failure or when the hazard rate of the system reaches or exceeds a certain level (control limit)*. Essentially, this is a control-limit policy with respect to the hazard rate. In our model, if we know the form of the baseline hazard function, and the link function, then for a certain state, (3.27) determines a unique threshold age for replacement because the hazard rate function is monotonic in time. Hence, the optimal replacement policy for our model can be uniquely defined by *n* threshold ages. Consider a system with a three-state *Z* process. As illustrated in Figure 3-1, the control limit d^* / K for the hazard rate fixes the planned replacement ages t_0, t_1, t_2 for state 0, 1, 2 respectively. Because the link function increases with the covariate state, we have $t_0 > t_1 > t_2$.

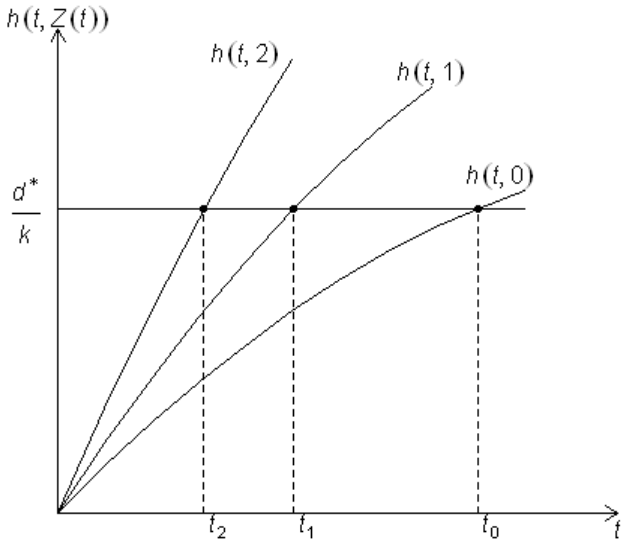

Figure 3-1 Replacement ages defined by the control limit.

We henceforth restrict our attention to the class of replacement policies in which a policy is composed of *n* threshold times for replacement, and we denote it as $\delta_{T_d} = \{t_0, t_1, ..., t_{n-1}\}$, $t_0 > t_1 > ... > t_{n-1}$, where t_i is the threshold age for replacement if the system is in state *i*. Obviously, the optimal policy in (3.27) falls within this class.

With the form of the optimal policy known from Theorem 1, and the iterative algorithm given in Proposition 3-1, there is still one barrier in the way of obtaining the optimal policy and cost for our model, which is the evaluation of (3.26), or how to compute the corresponding cost for a given stopping rule. An explicit expression for the objective function (3.23) in terms of the policy parameters $t_0, t_1, ..., t_{n-1}$ is necessary to overcome the barrier. We address this issue in the next section.

3.4 Explicit Expression of the Long-Run Average Cost

From (3.23), calculation of the objective involves evaluating the failure probability $\Pr(T_d \geq T)$, and *s*-expected time to replacement $E\left[\min\{T, T_d\}\right]$. For notational convenience, define $W_d = W(t_0, t_1, ..., t_{n-1}) = E\left[\min\{T, T_d\}\right]$ as the *s*-expected life of the system, and define $Q_d = Q(t_0, t_1, ..., t_{n-1}) = \Pr\left(T \leq T_d\right)$ as the probability of failure under policy $\delta_{T_d} = \{t_0, t_1, ..., t_{n-1}\}$. In what follows, we show that it is possible to explicitly represent W_d and Q_d as functions of $t_0, t_1, ..., t_{n-1}$ by conditioning on the time instants at which the system changes state; that is, on $S_0, S_1, ..., S_{n-2}$. For simplicity, we take the system with a three state covariate process as an illustration. The results generalize to situations with more states.

Assume the marginal pdfs of sojourn times X_0, and X_1 are $f_{X_0}(\cdot)$, and $f_{X_1}(\cdot)$ respectively. It follows that the pdf of S_0 is

$$g_0(s_0) = f_{X_0}(s_0). \tag{3.28}$$

Also, note that the event $\left[S_0 = s_0, S_1 = s_1\right]$ is equivalent to the event $\left[X_0 = s_0, X_1 = s_1 - s_0\right]$. Hence the joint pdf of S_0 and S_1 is

$$g_1(s_0, s_1) = f_{X_0}(s_0) f_{X_1}(s_1 - s_0),\ 0 < s_0 < s_1. \tag{3.29}$$

In accordance with the survivor function in (3.22), define the conditional Cdf of system failure time T as follows by conditioning on S_0 and S_1, where s_0 and s_1 are realizations of S_0 and S_1, respectively, and $s_0 < s_1$.

Let $F(t; s_0, s_1) \equiv \Pr(T \le t \mid S_0 = s_0, S_1 = s_1)$.

Then, for $t \le s_0$,

$$F(t; s_0, s_1) = F_0(t) \equiv 1 - \exp\left(-\psi(0)\int_0^t h_0(u)\,du\right).$$

For $s_1 > t > s_0$,

$$F(t; s_0, s_1) = F_1(t; s_0) \equiv 1 - \exp\left(-\psi(0)\int_0^{s_0} h_0(u)\,du - \psi(1)\int_{s_0}^t h_0(u)\,du\right).$$

For $t > s_1$,

$$F(t; s_0, s_1) = F_2(t; s_0, s_1) \equiv 1 - \exp\left(-\psi(0)\int_0^{s_0} h_0(u)\,du - \psi(1)\int_{s_0}^{s_1} h_0(u)\,du - \psi(2)\int_{s_1}^t h_0(u)\,du\right).$$

Again, conditioning on S_0 and S_1, there will be five different cases based on the relative positions among t_2, t_1, t_0 and s_0, s_1, as discussed below. Note that $t_2 < t_1 < t_0$, and $s_0 < s_1$. Under each case, the expressions of W_d and Q_d can be derived accordingly.

Let

$$W(t_0, t_1, t_2; s_0, s_1) \equiv E\left(\min\{T, T_d\} \mid S_0 = s_0, S_1 = s_1\right)$$

$$Q(t_0, t_1, t_2; s_0, s_1) \equiv \Pr(T \le T_d \mid S_0 = s_0, S_1 = s_1).$$

By the Law of Iterated Expectation [22],

$$W(t_0, t_1, t_2; s_0, s_1) = E\left(E\left(\min\{T, T_d\} \mid S_0, S_1, T\right) \mid S_0 = s_0, S_1 = s_1\right).$$

Case 0: If $s_0 > t_0$, then

$$\min\{T, T_d\} = \begin{cases} T & \text{if } T \le t_0 \\ t_0 & \text{if } T > t_0 \end{cases}$$

$$W(t_0, t_1, t_2; s_0, s_1) = W_0(t_0) \equiv \int_0^{t_0} t\,dF_0(t) + t_0\left[1 - F_0(t_0)\right]$$

$$Q(t_0, t_1, t_2; s_0, s_1) = Q_0(t_0) \equiv F_0(t_0).$$

Case 1: If $t_1 < s_0 < t_0$, then

$$\min\{T, T_d\} = \begin{cases} T & \text{if } T \le s_0 \\ s_0 & \text{if } T > s_0 \end{cases}$$

$$W(t_0, t_1, t_2; s_0, s_1) = W_1(s_0) \equiv \int_0^{s_0} t dF_0(t) + s_0 \left[1 - F_0(s_0)\right]$$

$$Q(t_0, t_1, t_2; s_0, s_1) = Q_1(s_0) \equiv F_0(s_0).$$

Case 2: If $s_0 < t_1$, $s_1 > t_1$ then

$$\min\{T, T_d\} = \begin{cases} T \text{ if } T \le t_1 \\ t_1 \text{ if } T > t_1 \end{cases}$$

$$W(t_0, t_1, t_2; s_0, s_1) = W_2(s_0, t_1) \equiv \int_0^{s_0} t dF_0(t) + \int_{s_0}^{t_1} t dF_1(s_0, t) + t_1 \left[1 - F_1(s_0, t_1)\right]$$

$$Q(t_0, t_1, t_2; s_0, s_1) = Q_2(s_0, t_1) \equiv F_1(s_0, t_1).$$

Case 3: If $s_0 < t_1$, $t_2 < s_1 < t_1$, then

$$\min\{T, T_d\} = \begin{cases} T \text{ if } T \le s_1 \\ s_1 \text{ if } T > s_1 \end{cases}$$

$$W(t_0, t_1, t_2; s_0, s_1) = W_3(s_0, s_1) \equiv \int_0^{s_0} t dF_0(t) + \int_{s_0}^{s_1} t dF_1(s_0, t) + s_1 \left[1 - F_1(s_0, s_1)\right]$$

$$Q(t_0, t_1, t_2; s_0, s_1) = Q_3(s_0, s_1) \equiv F_1(s_0, s_1).$$

Case 4: If $s_0 < t_1$, $s_1 < t_2$, then

$$\min\{T, T_d\} = \begin{cases} T \text{ if } T \le t_2 \\ t_2 \text{ if } T > t_2 \end{cases}$$

$$W(t_0, t_1, t_2; s_0, s_1) = W_4(s_0, s_1, t_2)$$
$$\equiv \int_0^{s_0} t dF_0(t) + \int_{s_0}^{s_1} t dF_1(s_0, t) + \int_{s_1}^{t_2} t dF_2(s_0, s_1, t) + t_2 \left[1 - F_2(s_0, s_1, t_2)\right]$$
$$Q(t_0, t_1, t_2; s_0, s_1) = Q_4(s_0, s_1, t_2) \equiv F_2(s_0, s_1, t_2).$$

With the above five cases at hand, by another application of the Law of Iterated Expectation,

$$W_d = E\left[E\left(\min\{T, T_d\} \mid S_0, S_1\right)\right] = \int_{t_0}^{\infty} W_0(t_0) g_0(s_0) ds_0 + \int_{t_1}^{t_0} W_1(s_0) g_0(s_0) ds_0$$
$$+ \int_{t_1}^{\infty} \int_0^{t_1} W_2(s_0, t_1) g_1(s_0, s_1) ds_0 ds_1 + \int_{t_2}^{t_1} \int_0^{s_1} W_3(s_0, s_1) g_1(s_0, s_1) ds_0 ds_1 + \int_0^{t_2} \int_0^{s_1} W_4(s_0, s_1, t_2) g_1(s_0, s_1) ds_0 ds_1$$

(3.30)

$$Q_d = E\left[P(T \le T_d \mid S_0, S_1)\right] = \int_{t_0}^{\infty} Q_0(t_0) g_0(s_0) ds_0 + \int_{t_1}^{t_0} Q_1(s_0) g_0(s_0) ds_0$$
$$+ \int_{t_1}^{\infty} \int_{0}^{t_1} Q_2(s_0, t_1) g_1(s_0, s_1) ds_0 ds_1 + \int_{t_2}^{t_1} \int_{0}^{s_1} Q_3(s_0, s_1) g_1(s_0, s_1) ds_0 ds_1 + \int_{0}^{t_2} \int_{0}^{s_1} Q_4(s_0, s_1, t_2) g_1(s_0, s_1) ds_0 ds_1$$

(3.31)

So far, we have obtained the integral expressions of W_d and Q_d in terms of the policy parameters t_0, t_1, t_2 for the system with a three-state covariate process. For a system with an *n*-state covariate process, there are $2n-1$ different cases. Thus the expression for W_d consists of $2n-1$ terms, each of which is an *n*-fold integral. The expression of Q_d is similar. Explicitly writing out the $(2n-1)$ *n*-fold integrals seems to be a formidable task. However, thanks to the connection between the *n* state model and the $(n+1)$ state model, this task is reduced to something tractable. In fact, for the $(n+1)$ state model, the expression for W_d has $2n+1$ cases, the first $2n-2$ cases of which are exactly the same as those of the W_d expression for the *n* state model, and the last three cases of which form a partition of the last case of the *n* state model by values of the new transition instant, S_n. Therefore, we can build the expressions of W_d and Q_d for an *n*-state covariate process by adding one state at a time. For comparison and illustration, we show the formulas for a system with a two-state covariate process in Appendix 3.A.

Based on the explicit expressions of W_d, Q_d, and Proposition 3-1, we describe the following iterative algorithm, which can be employed to find the optimal policy parameters and the optimal cost simultaneously.

Algorithm 3-1

1. Initialize the iteration counter $m = 0$. Choose an arbitrary replacement policy, and let d_0 equal the cost of the chosen policy.
2. For d_m, use (3.25) to find the threshold time t_i^m for replacement if the system state is in state i, i.e.,
$$t_i^m = \inf\left\{t \ge 0 : h_0(t)\psi(i) = d_m / K\right\}, i \in S. \quad (3.32)$$
3. Use the replacement policy $\delta_m = \{t_0^m, t_1^m, .., t_{n-1}^m\}$ obtained in step 2, (3.23), (3.30), and

(3.31) to update $d_{m+1} = \phi(\delta_m)$.

4. If $d_{m+1} = d_m$, stop with $d^* = d_{m+1}$, and $\delta^* = \{t_0^*, t_1^*, ..., t_{n-1}^*\} = \{t_0^m, t_1^m, ..., t_{n-1}^m\}$; otherwise, set $m \leftarrow m+1$, and go to step 2.

3.5 Numerical Example and Sensitivity Analysis

3.5.1 Numerical Example

To illustrate our model, and the procedure to construct the optimal policy, we consider a system with a three-state covariate process as a numerical example. In the following analysis, we assume that the functions that define the failure model, namely $h_0(t)$, $\psi(Z_t)$, and $f_{X_k}(x_k)$, are known, and their parameters are given (estimated). In practice, with historical monitoring data and lifetime data, the forms of those functions can be established either empirically, or through careful statistical analysis [23], [24]. The parameters of those functions can be estimated using the maximum likelihood method and its variants (to cope with the truncated and censored data), such as the one used in [16].

Assume the baseline hazard function is a Weibull hazard function given by

$$h_0(t) = \frac{bt^{b-1}}{a^b}$$

with $a = 1$ and $b = 2$; and suppose that $\psi(Z_t) = \exp(cZ_t)$ with $c = 2$. Assume $C = 5$, and $K = 25$. Because the forms of $h_0(t)$ and $\psi(Z_t)$ are predefined, the PH model here is parametric rather than semi-parametric, as described in [13].

Suppose the semi-Markov process Z has three states $\{0,1,2\}$, and the sojourn times X_0 and X_1 are *s*-independent identically distributed Weibull random variables with mean 1. The Weibull distribution is chosen here because it includes the exponential distribution as a special case, which allows convenient comparisons between systems with Markovian and semi-Markovian covariate processes. Assume the pdf of X_i is

$$f_{X_i}(x_i) = \frac{\beta x_i^{\beta-1}}{\eta^\beta} \exp\left[-\left(\frac{x_i}{\eta}\right)^\beta\right],\ 0 < x_i, i = 0,1$$

with $\beta = 1.5$, and $\eta = 1.1077$. It is not hard to check that the mean of X_i is approximately 1.

In Algorithm 3-1, we initialize $d_0 = (C + K)/E(T)$, which is the cost of the policy that replaces only at failure. The mean time to failure $E(T)$ could be obtained from (3.30) by setting $t_0 = t_1 = t_2 = \infty$. In this way, we find $E(T) = 0.6813$, and $d_0 = 44.0335$.

The complete results are shown in Table 3-1. The iterative algorithm converges after five iterations to the optimal average cost $d^* = 23.4364$. The algorithm was implemented in *Mathematica*®.

Table 3-1 Illustration of the Computation Procedure with Weibull(1.2089, 1.5) Sojourn Time

m	d_m	t_0^m	t_1^m	t_2^m	$W(t_0^m, t_1^m, t_2^m)$	$Q(t_0^m, t_1^m, t_2^m)$	$\phi(t_0^m, t_1^m, t_2^m)$
0	44.0335	0.8807	0.1192	0.016	0.5618	0.3846	26.0157
1	26.0157	0.5203	0.0704	0.0095	0.4248	0.1998	23.5262
2	23.5262	0.4705	0.0637	0.0086	0.3958	0.1710	23.4365
3	23.4365	0.4687	0.0634	0.0086	0.3947	0.1700	23.4364
4	23.4364	0.4687	0.0634	0.0086	0.3947	0.1700	23.4364

To study the effect of the parameters of the Weibull sojourn time, we varied the shape parameter β from 0.8 to 2, and changed the scale parameter η accordingly to ensure the same mean sojourn time. Table 3-2 shows the optimal replacement policies and costs for various Weibull sojourn time distributions (STD). We also include coefficients of variation (CV) of the distributions in Table 3-2 to gain more insight. One interesting observation is that the optimal cost increases with the CV of the STD, which is reasonable because in practice larger variability always tends to boost the cost.

Another notable observation is that different STDs lead to different optimal policies and costs, even if they all follow Weibull distributions, and have the same mean. This observation implies a pitfall if we always model the covariate process as Markovian. Suppose the true STD is Weibull(1.1077, 1.5). If we use the Markov model, then the best estimated STD is Weibull(1, 1); i.e., Exp(1), which would lead to a non-optimal replacement policy, and higher replacement cost. The cost errors for using policy parameters from the Markov model in other sojourn times are shown in Table 3-3. We can see that the relative error

becomes smaller as the CV of the true STD gets closer to 1. In this example, those errors are relatively small, which means that, when the STD of the covariate process is unknown, and hard to estimate, a Markov process might be a good candidate, and the investment for a good estimation of the STD would be of only marginal value. Besides, the Markov model could simplify the computation for the optimal policy because exponential STD would simplify the evaluation of the multiple integrals.

In the special case where the covariate process is Markovian; i.e., the STD is Weibull(1, 1), the computational procedure for periodic monitoring [18] can be used to approximate continuous monitoring by setting the monitoring interval to be very small. In that approach, recursion is needed for calculation of both $W(t_0,t_1,t_2)$ (s-expected life), and $Q(t_0,t_1,t_2)$ (failure probability), while the approach derived in this paper requires no recursion. This result gives the current approach a great computational advantage. For the example discussed here with a Weibull(1, 1) STD, the computational time to obtain the optimal policy and cost is 0.25 seconds on a computer with 1.83 GHz CPU, and 2GB main memory. However, if using periodic monitoring with an interval of 0.01 time units to approximate the continuous monitoring, the resulting policy is similar, but the computational time is 10.4 seconds, which is substantially longer. Based on this computational advantage, we suggest using the formulas in this paper to approximate the optimal policy under periodic monitoring when the monitoring interval is small, as well as to compute the exact optimal policy under continuous monitoring.

Table 3-2 Effect of Different Weibull Parameters on the Optimal Policy and Cost

Sojourn Time Distribution	CV	t_0	t_1	t_2	$W(t_0,t_1,t_2)$	$Q(t_0,t_1,t_2)$	d^*
Weibull(0.7900, 0.7)	1.4624	0.5293	0.0716	0.0097	0.3281	0.1473	26.4652
Weibull(0.8826, 0.8)	1.2605	0.5125	0.0694	0.0094	0.3428	0.1514	25.6249
Weibull(1, 1)	1	0.4913	0.0665	0.0090	0.3646	0.1582	24.5645
Weibull(1.1077, 1.5)	0.6790	0.4687	0.0634	0.0086	0.3947	0.1700	23.4364
Weibull(1.1284, 2)	0.5227	0.4609	0.0624	0.0084	0.4088	0.1769	23.0469

Table 3-3 Cost Errors for using Policy Parameters from a Markov model

Sojourn Time Distribution	CV	Absolute Error	Relative Error
Weibull(0.7900, 0.7)	1.4624	0.0453	0.171%
Weibull(0.8826, 0.8)	1.2605	0.0144	0.056%
Weibull(1.1077, 1.5)	0.6790	0.0185	0.079%
Weibull(1.1284, 2)	0.5227	0.0355	0.154%

Table 3-4 shows the optimal policies and costs for Lognormal STDs. Again, all of these distributions have the same mean, approximately equal to 1. Table 3-4 confirms the conclusion that a large CV for the STD has a harmful effect on the optimal cost. Besides, comparing similar cases in Table 3-2 and Table 3-4 suggests that, for the same CV, the Lognormal sojourn time leads to a lower optimal cost than the Weibull sojourn time.

Table 3-4 Optimal Policy and Cost when Sojourn time is Lognormal

Sojourn time Distribution	CV	t_0	t_1	t_2	$W(t_0,t_1,t_2)$	$Q(t_0,t_1,t_2)$	d^*
Lognor(-0.5, 1)	1.3108	0.4805	0.0650	0.0088	0.3691	0.1548	24.0264
Lognor(-0.3469, 0.83)	1	0.4680	0.0633	0.0086	0.3893	0.1645	23.4036
Lognor(-0.1922, 0.62)	0.6846	0.4585	0.0621	0.0084	0.4108	0.1770	22.9264
Lognor(-0.125, 0.5)	0.5329	0.4560	0.0617	0.0084	0.4192	0.1823	22.7990

3.5.2 Sensitivity Analysis

In the above numerical example, we assume all the model parameters are fixed. However, in practice, some of those parameters must be estimated from the historical data of the system. The quality of the estimates will directly affect the validity of the resulting replacement policy. In this subsection, we investigate how the variations in the model parameters impact the long-run average cost, and we assess the relative importance of model parameters through sensitivity analysis. In particular, we evaluate three input parameters, which are *a* and *b* in the baseline hazard function, $h_0(t) = bt^{b-1}/a^b$; and *c* in the link function, $\psi(Z_t) = \exp(cZ_t)$. (For simplicity, we assume the forms of $h_0(t)$ and $\psi(Z_t)$ are known, and all the other parameters are given and the same as in Subsection 3.5.1) We choose Weibull(1.1077, 1.5) as the STD for the *Z* process.

Assume the true parameter values are $a = 2, b = 2, c = 2$, and their estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$ each *s*-independently follow the distribution $N(2, 0.4)$. Performing the FAST sensitivity analysis method [25] with 1000 samples using SimLab [26], we get the FAST first-order indexes, as shown in Table 3-5. This index gives the expected reduction in the variance of the cost if an individual parameter is fixed. This table indicates that the scale parameter of the baseline hazard function, *a*, accounts for most of the variability in the output, and therefore is the most important of the three parameters. It implies that, if we can somehow reduce the variances of some input parameters' estimates by investing more, we should give parameter *a* the highest priority.

Notably, the conclusions reached by sensitivity analysis are case-specific, and should not be generalized if the model parameters are changed.

Table 3-5 FAST First-Order Indexes

Parameters	First-order indexes on cost
a	0.3329
b	0.1069
c	0.0383

3.6 Conclusion

In this paper, we studied the optimal replacement problem for general deteriorating systems. The aging and deterioration process is characterized by the proportional hazards model with a semi-Markovian covariate process, which we assume is under continuous monitoring. Allowing the covariate process to be semi-Markovian endows our method with great capability and flexibility to model real world situations. To minimize the long-run average cost per unit time, first we identified our model as a special case of Bergman's model [20], and determined that the optimal replacement policy of our model is of the control limit type with respect to the hazard function. Given that an optimal policy may be uniquely defined by a set of state-dependent threshold ages for replacement, an explicit expression for the objective function was derived in terms of those threshold ages by conditioning. Then the

iterative procedure developed by Bergman was customized for our model to find the optimal threshold ages.

A numerical example with n=3 covariate states illustrates the computational procedure, as well as the effects of different sojourn time distributions of the covariate process on the optimal policy and cost. The results show that larger variability in the sojourn time distributions (STD) tends to increase the cost of the optimal replacement policy. However, some numerical results show that, when the STD of the covariate process is difficult to estimate, viewing the process as a Markov process is not a bad option. Sensitivity analysis on an instance indicates that the variance of the scale parameter in the baseline hazard function accounts for most of the resulting variability in the cost, and therefore the scale parameter is of the most importance among the three chosen parameters.

Possible extensions of the research could be to 1) generalize the one-dimensional covariate to a multi-dimensional vector which would permit the Z process to evolve along multiple paths; 2) introduce uncertainty in the monitoring process, that is, the partial observation problem, to our current model; and 3) use a new failure model to relate the covariate information to system failure time distribution, such as an accelerated failure time model [23].

Appendix 3.A Formulas for System with a Two-State Covariate Process

For the system with a two-state covariate process, there will be only one time instant, S_0, at which the system changes states. In the following, we show how to explicitly represent the s-expected life of the system $W_d = W(t_0, t_1) = E\left[\min\{T, T_d\}\right]$, and the probability of failure $Q_d = Q(t_0, t_1) = \Pr\left(T \le T_d\right)$ under policy $\delta_d = \{t_0, t_1\}$ by conditioning on S_0.

Define the conditional Cdf of system failure time T as follows.

$$F(t; s_0) \equiv \Pr(T \le t \mid S_0 = s_0),$$

where s_0 is the realization of S_0.

Then for $t \le s_0$,

$$F(t;s_0) = F_0(t) \equiv 1 - \exp\left(-\psi(0)\int_0^t h_0(u)\,du\right).$$

For $t > s_0$,

$$F(t;s_0) = F_1(t;s_0) \equiv 1 - \exp\left(-\psi(0)\int_0^{s_0} h_0(u)\,du - \psi(1)\int_{s_0}^t h_0(u)\,du\right).$$

Let

$$W(t_0,t_1;s_0) \equiv E\left(\min\{T,T_d\} \mid S_0 = s_0\right)$$

$$Q(t_0,t_1;s_0) \equiv \Pr(T \le T_d \mid S_0 = s_0).$$

By the Law of Iterated Expectation [22],

$$W(t_0,t_1;s_0) = E\left(E\left(\min\{T,T_d\} \mid S_0,T\right) \mid S_0 = s_0\right).$$

There will be three cases.

Case 0: If $s_0 > t_0$, then

$$\min\{T,T_d\} = \begin{cases} T & \text{if } T \le t_0 \\ t_0 & \text{if } T > t_0 \end{cases}$$

$$W(t_0,t_1;s_0) = W_0(t_0) \equiv \int_0^{t_0} t\,dF_0(t) + t_0\left[1 - F_0(t_0)\right]$$

$$Q(t_0,t_1;s_0) = Q_0(t_0) \equiv F_0(t_0).$$

Case 1: If $t_1 < s_0 < t_0$, then

$$\min\{T,T_d\} = \begin{cases} T & \text{if } T \le s_0 \\ s_0 & \text{if } T > s_0 \end{cases}$$

$$W(t_0,t_1;s_0) = W_1(s_0) \equiv \int_0^{s_0} t\,dF_0(t) + s_0\left[1 - F_0(s_0)\right]$$

$$Q(t_0,t_1;s_0) = Q_1(s_0) \equiv F_0(s_0).$$

Case 2: If $s_0 < t_1$, then

$$\min\{T,T_d\} = \begin{cases} T & \text{if } T \le t_1 \\ t_1 & \text{if } T > t_1 \end{cases}$$

$$W(t_0,t_1;s_0) = W_2(s_0,t_1) \equiv \int_0^{s_0} t\,dF_0(t) + \int_{s_0}^{t_1} t\,dF_1(s_0,t) + t_1\left[1 - F_1(s_0,t_1)\right]$$

$$Q(t_0,t_1;s_0) = Q_2(s_0,t_1) \equiv F_1(s_0,t_1).$$

Then by another application of the Law of Iterated Expectation,

$$W_d = E\left[E\left(\min\{T, T_d\} \mid S_0\right)\right] = \int_{t_0}^{\infty} W_0(t_0) g_0(s_0) ds_0 + \int_{t_1}^{t_0} W_1(s_0) g_0(s_0) ds_0 + \int_0^{t_1} W_2(s_0, t_1) g(s_0) ds_0 ,$$

$$Q_d = E\left[P(T \le T_d \mid S_0)\right] = \int_{t_0}^{\infty} Q_0(t_0) g_0(s_0) ds_0 + \int_{t_1}^{t_0} Q_1(s_0) g_0(s_0) ds_0 + \int_0^{t_1} Q_2(s_0, t_1) g_0(s_0) ds_0 .$$

Comparison with (3.30) and (3.31) shows the recursive nature of these expressions.

Acknowledgements

This work was supported by the National Science Foundation under grant CNS-0540293.

References

[1] M. Wang, A. Vandermaar, and K. Srivastava, "Review of condition assessment of power transformers in service," *IEEE Electrical Insulation Magazine*, vol. 18, no. 6, 2002.

[2] B. Augenstein, "Outside experts monitor status of key transformers," *Transmission & Distribution World*, May 2003. http://tdworld.com/mag/power_outside_experts_monitor/ (viewed September 2010).

[3] W. Q. Meeker, "Trends in the statistical assessment of reliability," tech. rep., Iowa State University, 2009.

[4] H. Mine and H. Kawai, "An optimal inspection and replacement policy," *IEEE Transactions on Reliability*, vol. 24, pp. 305–309, 1975.

[5] H. Mine and H. Kawai, "An optimal inspection and replacement policy of a deteriorating system," *Journal of Operations Research Society of Japan*, vol. 25, pp. 1–15, 1982.

[6] N. Douer and U. Yechiali, "Optimal repair and replacement in Markovian systems," *Communications in Statistics – Stochastic Models*, vol. 10, pp. 253–270, 1994.

[7] C. T. Lam and R. H. Yeh, "Optimal replacement policies for multi-state deteriorating systems," *Naval Research Logistics*, vol. 41, no. 33, pp. 303–315, 1994.

[8] D. Chen and K. S. Trivedi, "Optimization for condition-based maintenance with semi-Markov decision process," *Reliability Engineering & System Safety*, vol. 90, no. 1, pp. 25–29, 2005.

[9] A. Pavitsos and E. G. Kyriakidis, “Markov decision models for the optimal maintenance of a production unit with an upstream buffer,” *Comput. Oper. Res.*, vol. 36, no. 6, pp. 1993–2006, 2009.

[10] A. Barros, C. Berenguer, and A. Grall, “Optimization of replacement times using imperfect monitoring information,” *IEEE Transactions on Reliability*, vol. 52, no. 4, 2003.

[11] J. Ivy and S. Pollock, “Marginally monotonic maintenance policies for a multi-state deteriorating machine with probabilistic monitoring, and silent failures,” *IEEE Transactions on Reliability*, vol. 54, no. 3, 2005.

[12] R. Toscano and P. Lyonnet, “On-line reliability prediction via dynamic failure rate model,” *IEEE Transactions on Reliability*, vol. 57, no. 3, 2008.

[13] D. R. Cox and D. Oakes, *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[14] A. K. S. Jardine and A. H. C. Tsang, *Maintenance, Replacement & Reliability: Theory and Applications*. Taylor & Francis Group, 2006.

[15] V. Makis and A. K. S. Jardine, “Optimal replacement in the proportional hazards model,” *INFOR*, vol. 30, no. 1, pp. 172–183, 1992.

[16] D. Banjevic, A. K. S. Jardine, V. Makis, and M. Ennis, “A control-limit policy and software for condition-based maintenance optimization,” *INFOR*, vol. 39, pp. 32–50, 2001.

[17] S. Ghasemi, S. Yacout, and M. S. Ouali, “Optimal condition based maintenance with imperfect information and the proportional hazards model,” *International Journal of Production Research*, vol. 45, no. 4, pp. 989–1012, 2007.

[18] X. Wu and S. M. Ryan, “Value of condition monitoring for optimal replacement in the proportional hazards model with continuous time degradation,” *IIE Transactions*, vol. 42, pp. 553–563, 2010.

[19] D. J. Woodcock, “Maintenance strategies for the new millennium – condition appraisal of power transformers,” 2000.

[20] B. Bergman, “Optimal replacement under a general failure model,” *Adv. in Appl. Prob.*, vol. 10, no. 2, pp. 431–541, 1978.

[21] S. M. Ross, *Introduction to Probability Models*. San Diego, CA: Academic Press, 8th ed., 2003.

[22] P. Billingsley, *Probability and Measure*. New York, NY: John Wiley & Sons, Inc., 3rd ed., 1995.

[23] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data*. New York: Wiley, 1998.

[24] Y. Hong, W. Q. Meeker, and J. McCalley, "Prediction of remaining life of power transformers based on left truncated and right censored lifetime data," *Annals of Applied Statistics*, vol. 3, pp. 857–879, 2009.

[25] A. Saltelli, S. Tarantola, F. Campolongo, and R. Marco, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. New York: John Wiley & Sons, 2004.

[26] European Union, "Simlab – sensitivity analysis," 2009. http://simlab.jrc.ec.europa.eu/ (viewed September 2010).

CHAPTER 4 JOINT OPTIMIZATION OF ASSET AND INVENTORY MANAGEMENT IN THE A PRODUCT-SERVICE SYSTEM

A paper submitted to *IIE Transactions*

Xiang Wu and Sarah M. Ryan

Abstract

This article proposes an integrated model of the asset management decisions for a fleet of products and the inventory management decisions for a closed-loop supply chain in the context of a product-service system, in which the two types of decisions are closely coupled. A joint optimization technique is developed to obtain the parameters of the operational policy for the integrated model that minimize the long run average cost per unit time. A numerical example is provided to illustrate the computational procedures. In addition, the effect of a simplifying assumption that the replaced products have no quality difference is evaluated and the results suggest that as long as the quality difference between the preventively replaced products and failure replaced products is not too big, the simplification to treat them as one category is reasonable.

Keywords: Joint optimization, product-service system, preventive maintenance, inventory management, closed-loop supply chain

4.1 Introduction

A product-service system (PSS), or servicizing[3], is a strategy in which producers provide the use as well as the maintenance of products while retaining ownership. Prospective customers who become the clients pay fees for receiving the services or functions of products rather than purchasing them, and so are free of the risk, responsibility and cost burdens that are commonly associated with ownership. Since the introduction of this attractive concept in 1999 (Goedkoop et al., 1999, White et al., 1999), a diverse range of PSS examples in the literature have demonstrated its economic success, but most have tended to emphasize its significant environmental benefits and social gains (Luiten et al., 2001, Manzini et al., 2001). Although a variety of tools and methodologies have been developed for designing a servicizing system, such as those in Manzini et al. (2001), Maxwell and van der Vorst (2003), and Van Halen et al. (2004), how to effectively structure an organization to be competent at designing, making and delivering PSS is still difficult (Baines and Lightfoot, 2007). Most literature in this area provides qualitative description and analysis of servicizing, and there is a lack of in-depth and rigorous research to develop models, methods and theories, to assess the implications of competitiveness, and to help manufacturers configure their products, technologies, operations, and supply chain (Baines and Lightfoot, 2007).

The motivation for this research is to improve the economic viability of PSS. Rather than examining the benefits of PSS in a case study, we analyze and model the operation of the PSS and identify optimal parameters of a policy to minimize the overall cost in the long run.

Providing product-based services requires the producer to extend its responsibility for the product both during and after the use phase. The service contracts frequently include replacement of the initial machines with newer or better ones, and the machines coming off lease are remanufactured extensively (Thierry et al., 1995). Service providers must balance the cost of building in durability and reusability against the lifecycle cost savings, choose when to take old products out of service, and decide whether to remanufacture them or to replace them with newly manufactured products. Servicizing motivates the use of condition monitoring; i.e., using sensors, information and communication technology to increase

[3] The view of servicizing is quite similar to PSS. In this paper, I will not distinguish the two concepts.

visibility of the product's performance and condition in the field, so as to improve asset utilization and make better maintenance decisions (Baines and Lightfoot, 2007). Under servicizing, the remanufacturing facilities frequently operate together with a manufacturing plant to satisfy the demand. Such systems are known as hybrid manufacturing and remanufacturing systems, and involve both forward and reverse flows of products.

For the service paradigm to be viable from the provider's perspective, the fee for service must allow for a profit margin over the cost of providing the service. The cost of service provision depends largely on the ability to manage and maintain products effectively in a closed-loop system. In particular, manufacturers who servicize must engage in reverse as well as forward logistics; and in addition, they must make maintenance decisions for their products. Unlike the common closed-loop supply chain for sold products, a distinctive feature of the closed-loop supply chain in PSS is that the demands are driven by maintenance actions on the products and/or a capacity expansion requirement, and the returns are generated by out-of-service products, replaced either preventively or due to failure. In other words, the demands and returns are controllable by the servicizing manufacturers via maintenance decisions, and the cost of replacement is affected by the inventory management decisions. Therefore, the maintenance decisions are closely coupled with the inventory management decisions of this closed-loop supply chain. This coupling makes the decision making under servicizing significantly more complicated than that under traditional product sales.

Maintenance policies for deteriorating systems have been studied extensively for decades (Aven and Bergman, 1986, Lam and Yeh, 1994, Liu et al., 2010, Giorgio et al., 2011). The recent research effort has been focused on the problem of optimal replacement when some condition information about the system is available, such as temperature, humidity, vibration levels, or the amount of metal particles in a lubricant, which is often the case in PSS (Banjevic et al., 2001, Ghasemi et al., 2007, Kharoufeh et al., 2010, Wu and Ryan, 2010). A rapidly growing body of research in the operations management of closed-loop supply chains recognizes and tries to mitigate the complexities of managing the supply chain involving remanufacturable products under traditional product sales (Fleischmann et al., 1997, Guide, 2000, Aras et al., 2004, Guide and Van Wassenhove, 2009).

However, little research has been done to consider the joint optimization of the maintenance policy and the closed-loop supply chain inventory management, to develop optimal decisions in the context of PSS. Some relevant work appears in the context of production inventory systems. For example, Das and Sarkar (1999) considered the optimal maintenance policies for a production inventory system where inventory is controlled according to an (S, s) policy. Rezg et al. (2004) studied the joint optimization problem of preventive maintenance and inventory control in a production line using simulation, and proposed a methodology combining simulation with genetic algorithms to obtain the optimal policy. In those cases, the maintenance is applied to the machines in the production line, rather than the service products in the fleet under PSS. Thus, their problems differ in nature from the one in PSS.

More relevant work in the existing literature investigates the joint optimization of maintenance and inventory policies for deteriorating systems with spare-part inventory. In particular, Armstrong and Atkins (1996) examined the age replacement and ordering decisions for a system subject to random failure and with room for only one spare in stock, and several extensions have been made to generalize the cost terms and the order lead time in their subsequent paper (Armstrong and Atkins, 1998). Brezavscek and Hudoklin (2003) considered the problem of joint optimization of block replacement and periodic review spare-provisioning policy for deteriorating systems to minimize the expected total cost per unit time. Ilgin and Tunali (2007) proposed a simulation optimization approach using genetic algorithms for the joint optimization of preventive maintenance and spare provisioning policies of a deteriorating system. Still, those studies differ fundamentally from the one we conduct in this paper, because they do not involve the production process of the spare parts.

In this work, we present an integrated model that takes into account both the replacement decisions and the inventory management decisions in the context of a product-service system to minimize the total cost per unit time. For maintenance, we consider a condition-based replacement policy that uses the proportional hazards model (PHM) with a semi-Markovian covariate process to model the degradation of the products (Wu and Ryan, 2011). For inventory management, a continuous review base stock policy is adopted due to its easy implementation and proven effectiveness in practice. Identifying and formulating the

couplings between asset and inventory management in this context, we develop an optimization technique to obtain the optimal parameters for the two policies simultaneously in the integrated model.

This paper is organized as follows. Section 4.3 presents the development and mathematical formulation of the final integrated model. In section 4.4, an optimization technique is developed and a two-step algorithm is presented to obtain the optimal policy parameters and cost. This is followed by a numerical example in Section 4.5 to illustrate the computational procedures of the optimization algorithm. Section 4.6 revisits the single return category assumption and evaluates its impact on the optimal cost by comparison with the analysis of a system with two categories of returns. Section 4.7 concludes with a discussion of future research directions.

4.2 System Description

We assume the service provider has a fleet of N identical products in service. The objective is to develop an operational policy for the PSS to keep every product in the fleet in working condition at all times with minimum cost. The products deteriorate with age and operation, and are subject to failure. When a product is preventively replaced or fails, it is collected for remanufacturing and replaced by a new product. The output of the remanufacturing facility may not be able to fulfill all the demand for new products. We assume a manufacturing facility exists with sufficient capacity to cover any unsatisfied demand.

The PSS consists of two subsystems: a service subsystem (SS) and a remanufacturing subsystem (RS) which is supplemented as needed by the manufacturing facility. The service subsystem employs products to provide service to clients and sends replaced products to the remanufacturing subsystem. The (hybrid) remanufacturing subsystem satisfies the demands of the service subsystem for replacement products through remanufacturing or manufacturing.

The flow of products through the whole system is depicted in Figure 4-1. In the SS, the operational conditions of the products are continuously monitored and a condition-based replacement policy is applied to each product independently. The demand of the

remanufacturing system is driven by the replacement of products in the fleet. When a product is taken out of service due to preventive replacement or failure, it is replaced immediately with either a remanufactured product or a newly manufactured product.

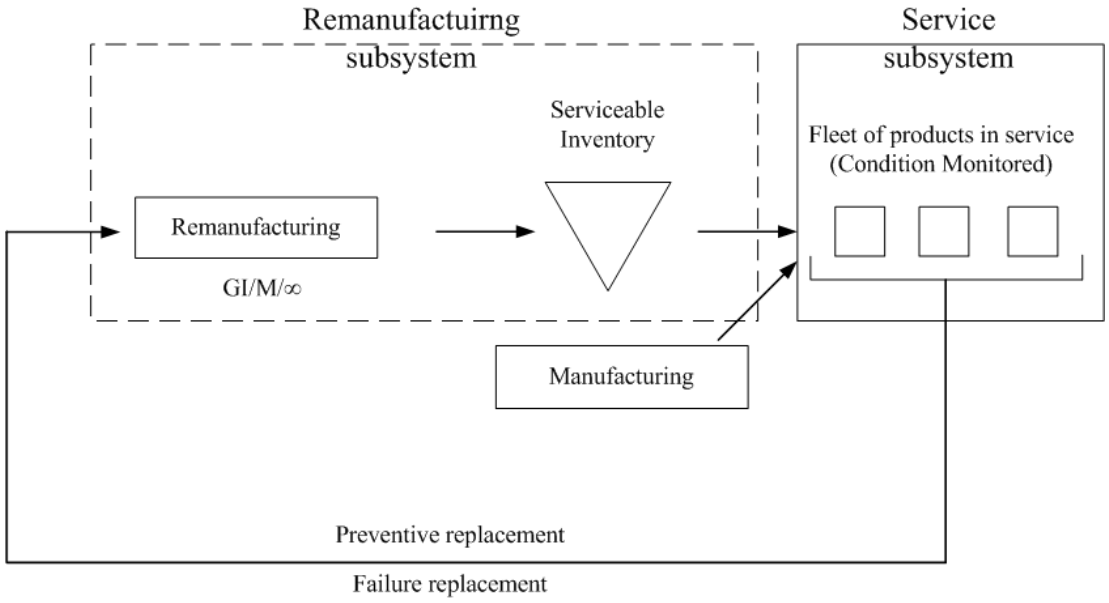

Figure 4-1 Product flow through the whole system

The RS is a remanufacturing facility that replenishes serviceable inventory. The replaced products directly go to the remanufacturing process if needed to maintain a base stock level; otherwise, they are discarded to save storage costs. We focus on the remanufacturing facility and do not represent the manufacturing plant in detail. Priority is given to remanufactured products when satisfying demand. Based on the conventional wisdom that remanufacturing is cheaper than manufacturing, newly manufactured products are viable only when the serviceable inventory is unable to fulfill the demand (i.e., is empty). We assume manufactured products are available whenever necessary.

The goal of the study is to investigate the replacement problem of the SS and the inventory management of the RS jointly in the context of PSS. Considering the coupling between two subsystems, an integrated model is built to address the uncertainties residing not only in the replacement problem but also in the inventory management, and a joint operational policy is developed to minimize the long-run average cost incurred in the whole system per unit time. In what follows, we shall first introduce the replacement policy for the

SS and the inventory management policy for the RS respectively, and then present the integrated model. First, we summarize the notations and assumptions used in this paper here.

4.2.1 Notation

Input parameters

N : Number of products in the fleet.

C_1 : The cost of preventive replacement with a remanufactured product in the SS, which is also the unit remanufacture (production) cost in the RS; $C_1 > 0$.

C_2 : The cost of preventive replacement with a manufactured product in the SS, which is also the aggregate acquisition cost for a newly manufactured product in the RS; $C_2 > C_1$.

K : The additional cost for a failure replacement; $K > 0$.

$Z = \{Z_t, t \geq 0\}$: A continuous semi-Markov process with a finite state space $S = \{0,1,...,n-1\}$ and $Z_0 = 0$, which depicts the evolution of the working condition of a product.

$h_0(t)$: The baseline hazard rate, which depends only on the age of the product.

$\Psi(\bullet)$: A link function; $\Psi : S \mapsto \Re$.

T : The time to failure of the product.

μ : Processing rate for remanufacturing.

h_S : Unit serviceable inventory holding cost.

h_W : Unit remanufacturing work in process (WIP) holding cost.

Internal variables

$\delta = \{t_0, t_1, \cdots, t_{n-1}\}$: A replacement policy which replaces at failure or at age t_i when in state i, whichever occurs first.

$M(\delta)$: The expected length of a replacement cycle under policy δ.

$Q(\delta)$: The probability of failure under policy δ.

$I(t)$: The number of products in the serviceable inventory at time t.

$W(t)$: The number of products in WIP at time t.

c : Base stock level of the serviceable inventory position.

p_L : The proportion of time that the serviceable inventory is empty.

Output variables

δ^* : Optimal replacement policy

c^* : Optimal base stock level

4.2.2 Assumptions

1 The service provider has a large fleet of identical products in service and maintains an inventory of serviceable products to satisfy the demand for replacements.

2 Manufactured and remanufactured products are perfectly substitutable; that is, remanufactured products are considered as good as new.

3 Setup cost for remanufacturing is negligible and there is no holding cost associated with the remanufacturable inventory.

4 Remanufacturing capacity is unlimited and the time required to remanufacture a replaced product is exponentially distributed with rate μ.

5 The newly manufactured products are always available and there is no lead time for acquiring one.

6 We consider the replaced products as one category, whether they are replaced preventively or due to failure. In section 4.6, we will re-evaluate this assumption.

7 The baseline hazard rate, $h_0(t)$, is strictly increasing with the product age and unbounded as the age approaches infinity; that is, the product deteriorates with time. In addition, $h_0(0) = 0$.

8 The covariate process Z changes state according to a pure birth process; i.e., whenever a transition occurs, the state of the process always increases by one, and state $n-1$ is absorbing.

9 The link function, $\Psi(\cdot)$, is non-decreasing with $\Psi(0) = 1$.

10 The fleet of products must be kept in working order at all times. Replacement is instantaneous

4.3 Model Development and Formulation

4.3.1 Replacement Policy for the Service Subsystem

In a PSS, the service provider retains ownership and maintains direct access to its products. This allows it to continuously collect data on the condition of products in service using condition monitoring technologies. Such data can help the service provider to improve the performance of products, lower failure probability, improve asset utilization and so reduce the total cost. For systems under continuous monitoring, a condition-based maintenance policy is natural.

Assume that replacement is the only maintenance option in our PSS setting. The condition-based replacement policy developed in Wu and Ryan (2011) well suits the service subsystem, where the PHM (Cox, 1984) is employed to account for the impact of dynamic working conditions on the failure process of the system. Herein we adopt the policy described in Wu and Ryan (2011) as the replacement policy for the SS.

Because the replacement policy is applied to each product independently, we first consider the replacement policy for a single product.

We assume that $Z = \{Z_t, t \geq 0\}$ is a continuous-time semi-Markov covariate process that depicts the evolution of the working condition of the product, and is under continuous monitoring. Under the proportional hazards model, the hazard rate of the product at time t is expressed as

$$h(t) \equiv h_0(t)\Psi(Z_t), \quad t \geq 0 \tag{4.33}$$

Denote the replacement policy as $\delta = \{t_0, t_1, \cdots, t_{n-1}\}$, $t_0 \geq t_1 \geq \cdots \geq t_{n-1} \geq 0$, where t_i is the threshold age for replacement if the covariate process of the product is in state i. According to renewal theory (Ross, 2003), the long run average replacement cost per unit time for a single product can be expressed as the ratio of the expected cost per replacement cycle to the expected length of a replacement cycle, which is given by

$$\phi_R = \frac{(C_1 + KQ(\delta))(1 - p_L) + (C_2 + KQ(\delta))p_L}{M(\delta)} = \frac{C_1 + (C_2 - C_1)p_L + KQ(\delta)}{M(\delta)} \tag{4.34}$$

Here p_L is the proportion of products in the fleet replaced with manufactured products, which will be discussed further in the context of the remanufacturing subsystem. The explicit expressions for the expected length of a replacement cycle, $M(\delta)$, and the failure probability, $Q(\delta)$, in terms of $t_0, t_1, \cdots, t_{n-1}$ given $Z(t)$, $h_0(t)$ and $\Psi(\cdot)$ can be found using the method described in Wu and Ryan (2011). We detail those expressions for a three-state Z process and their partial derivatives with respect to t_i in Appendix 4.A for convenience. Obviously, $M(\delta)$ is an increasing function of the threshold age for replacement t_i, $\forall i$.

4.3.2 Inventory Policy for the Remanufacturing Subsystem

For systems involving remanufacturing, two inventory control strategies are generally applied: "push" and "pull". Under the push strategy, the returned products are batched and pushed into the remanufacturing process as soon as the remanufacturable inventory has sufficient products. Under the pull strategy, the timing of the remanufacturing process depends on the demands as well as inventory positions. Van der Laan et al. (1999) shows that pull control is preferable if the holding cost in remanufacturable inventory is lower than the holding cost in the serviceable inventory, which is true in most practical situations.

Based on above findings, the inventory in the remanufacturing subsystem is assumed to be managed by a continuous review base stock policy. This policy aims at keeping the serviceable inventory position at a base stock level c at all times, which is achieved by pulling returned products into the remanufacturing process each time a demand is served from the serviceable inventory. The serviceable inventory position at time t includes the on-hand serviceable inventory $I(t)$ and the work in process (WIP) of remanufacturing $W(t)$. Thus we have

$$I(t) + W(t) = c \qquad \forall t \geq 0. \tag{4.35}$$

The policy is easy to implement and is efficient when the setup cost for remanufacturing is negligible, which we assume is true in our PSS setting.

Under the continuous review base stock policy, the remanufacturing subsystem is a pull system. An execution flowchart is shown in Figure 4-2. Priority is given to remanufactured

products when satisfying demand. Following the flowchart, we can see that p_L defined in Section 4.3.1 equals the probability that the serviceable inventory is empty; i.e.,

$$p_L = P(I(t) = 0)$$

The cost C_1 defined in Section 4.3.1 is equivalent to the unit remanufacturing cost, and C_2 is equivalent to the unit acquisition cost for a manufactured product, which is incurred every time we resort to manufacturing.

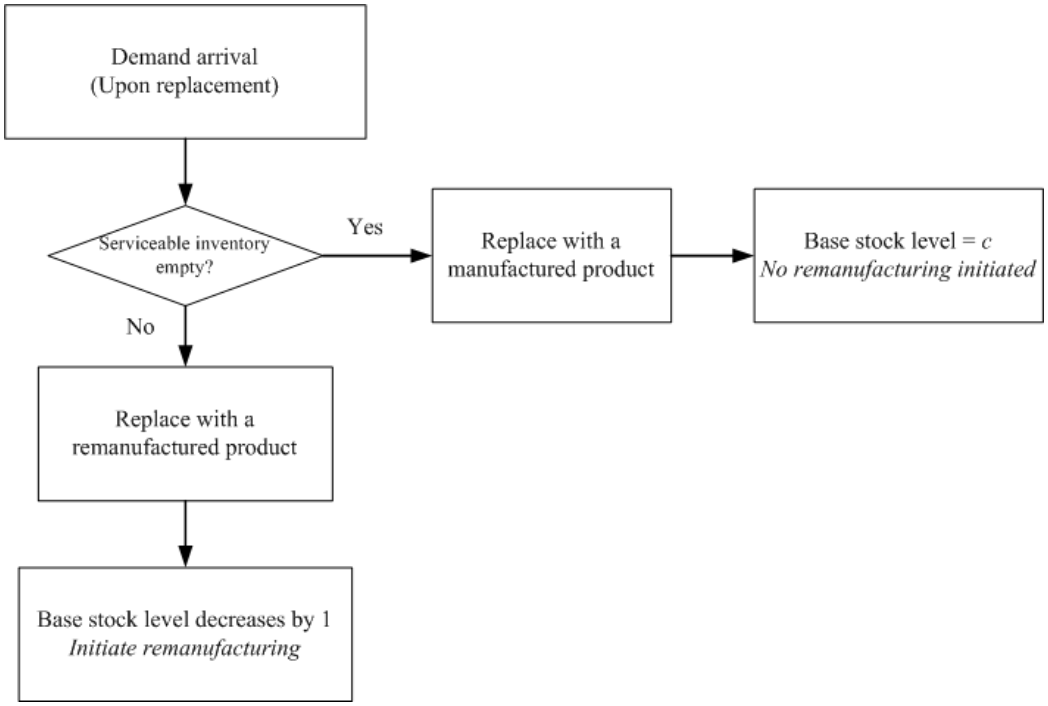

Figure 4-2 Flowchart of the remanufacturing subsystem

For the serviceable inventory and remanufacturing WIP, we adopt a similar holding cost structure to that in Aras et al. (2004). The unit holding cost rate for serviceable inventory is $h_S = h + \alpha C_1$, and the unit holding cost rate for remanufacturing WIP is $h_W = h + \beta\alpha C_1$ ($\beta < 1$). Here, h denotes the basic holding cost and α denotes the opportunity cost of capital. WIP is considered to have approximately $100\beta\%$ value added and the serviceable inventory has all the value added.

In addition, we assume there is no capacity limitation on the remanufacturing process, so it could be modeled as an infinite-server station. The time for remanufacturing is highly variable due to various conditions of the replaced products. Thus the service time of each

server is assumed to be exponentially distributed with rate μ. In fact, since the WIP in the remanufacturing subsystem is bounded by the base stock level c, only c servers are needed to avoid blocking in the remanufacturing process, and the subsystem can achieve steady state. So the loss probability p_L is constant over time

$$p_L = P(I = 0) \tag{4.36}$$

and the long run average cost incurred in the remanufacturing subsystem is given by

$$\phi_I = h_S E(I) + h_W E(W) \tag{4.37}$$

The t in the notations $I(t)$ and $W(t)$ has been suppressed. Only inventory costs are considered for the RS because we already account for the costs C_1 and C_2 in the SS.

We consider the returns as a single category, regardless of whether they are preventively replaced or replaced due to failure. We do not differentiate the returned products in terms of inventory cost, processing time and cost. Therefore the remanufacturing node together with the serviceable inventory node can be modeled as a single-stage produce-to-stock system with a single product type.

Examining our system carefully, we find that the inventory is virtually controlled by a target-level production authorization mechanism with lost sales as discussed in Buzacott and Shanthikumar (1993). In our case, the target-level is c. Production authorization is transmitted to the remanufacturing facility when the inventory position falls by one. A "lost sale" occurs when the serviceable inventory is empty, in which situation we resort to manufactured products and no new remanufacturing is authorized. According to Buzacott and Shanthikumar (1993), the performance of this produce-to-stock system may be obtained from the analysis of a fictitious $G/M/c/c$ queue, also known as $G/M/c$ loss system. The correspondence between our system and the fictitious system is as follows:

- The demand process to the RS is the arrival process to the fictitious queue.
- The probability that a demand is satisfied by manufacturing, p_L, is the loss probability of the fictitious queue.
- The WIP of the RS is the number of products in the fictitious queue.

The demand process of the RS is generated from the replacements of products in the SS. Since the replacement policy is applied to each product independently, the replacement flow of each individual product is a renewal process. The demand process, which is a pool of N such renewal processes, is called a superposed renewal process (SRP) in the literature. In general, if the number of products in a service fleet, N, is sufficiently large, then the superposed renewal process can be approximated by a Poisson process with rate λ (Cinlar and Lewis, 1972). Then the fictitious queue may be approximated by a $M/M/c$ loss system, also known as the Erlang loss system.

For each product, the renewal rate $r = \frac{1}{M(\delta)}$. Then the overall arrival rate is

$$\lambda = Nr = \frac{N}{M(\delta)} \tag{4.38}$$

Let L be the average number of products in the $M/M/c$ loss system and p_n be the steady-state probability that there are n products in the queue. From the performance of $M/M/c$ loss system in steady state, we have

$$p_0 = 1/\sum_{k=0}^{c} \frac{(\lambda/\mu)^k}{k!}$$

$$p_n = \frac{(\lambda/\mu)^n}{n!} p_0, \quad n = 1, 2, ., c$$

$$L = \frac{\lambda}{\mu}(1 - p_c)$$

For our system,

$$E(W) = L \tag{4.39}$$

$$E(I) = c - L \tag{4.40}$$

From equations (4.37)-(4.40),

$$\phi_I = h_S c + (h_W - h_S)\frac{\lambda}{\mu}(1 - p_L) \tag{4.41}$$

where the probability that a demand must be satisfied by manufacturing is

$$p_L \equiv p_L(c, \delta) = \lim_{t \to \infty} P(I(t) = 0) = p_c = \frac{(\lambda/\mu)^c}{c!\sum_{k=0}^{c} \frac{(\lambda/\mu)^k}{k!}} \tag{4.42}$$

with $\lambda = N / M(\delta)$.

We state two important properties of p_L as below.

Lemma 4-1 $\frac{\partial p_L}{\partial M} < 0$.

Proof.

The loss probability p_L is a increasing function of the arrival rate λ, and thus a decreasing function of $M(\delta)$. Therefore $\frac{\partial p_L}{\partial M} < 0$. ■

Lemma 4-2 $\frac{\partial p_L}{\partial M}$ *is increasing in its parameter* $t_i, \forall i$.

Proof.

According to Proposition 3 of Harel (1990), for fixed number of servers and fixed arrival rate, the Erlang loss formula is strictly convex in service rate. The symmetric positions of μ and $M(\delta)$ in the loss formula (4.42) imply that p_L is strictly convex in $M(\delta)$ for fixed c and μ. Thus $\frac{\partial^2 p_L}{\partial M^2} > 0$, which means $\frac{\partial p_L}{\partial M}$ is increasing in $M(\delta)$. And since $M(\delta)$ is an increasing function of t_i, $\frac{\partial p_L}{\partial M}$ is increasing in t_i, $\forall i$. ■

4.3.3 Integrated Model

The service subsystem (SS) and the remanufacturing subsystem (RS) discussed above are closely coupled in terms of the demands and returns. The RS satisfies the demand of the SS, while the SS generates returns to the RS. A distinctive feature of this closed-loop system is that the demands and the returns are generated simultaneously.

In particular, the decision making process couples the two subsystem. The decision variable in the SS, the replacement policy δ, affects the demand and return flows of the supply chain in the RS and, thus, affects the average cost incurred in the RS. On the other hand, the base stock level c has a direct impact on p_L, the proportion of products replaced with manufactured products and thus, influences the average cost incurred in the SS.

To account for the coupling, we must optimize the decision variables of the two subsystems simultaneously. Treating them separately would lead to an inferior solution.

Integrating the costs of subsystems, the total cost per unit time for the whole system can be expressed as

$$\phi(c,\delta) = \phi_I + N\phi_R = h_S c + N\frac{C_1 + (h_W - h_S)\mu^{-1} + KQ(\delta) + [C_2 - C_1 - (h_W - h_S)\mu^{-1}]p_L}{M(\delta)} \quad (4.43)$$

which incorporates the production cost for remanufactured products (C_1 per unit), the acquisition cost for manufactured products (C_2 per unit), additional cost for failure replacement (K per unit) and the inventory holding costs (h_S per unit per unit time for serviceable inventory and h_W per unit per unit time for WIP).

4.4 Optimization Technique

Our objective function is the total cost per unit time given by (4.43). The decision variables are the parameters of the replacement policy $\delta = \{t_0, t_1, \cdots, t_{n-1}\}$ and the base stock level c, where $t_0 \geq t_1 \geq \cdots \geq t_{n-1} \geq 0$ and c is a positive integer. The key variables are p_L, expressed in (4.42); and $M(\delta)$ and $Q(\delta)$, whose expressions can be obtained from Appendix 4.A.

The objective function is a complicated function that appears to lack "nice" structure (such as convexity) and closed-form solutions are hard to achieve. To obtain the global minimum, we resort to a special optimization method -- the lambda minimization technique (Aven and Bergman, 1986) which is summarized in Appendix 4.B. For simplicity, we consider a three-state covariate process Z to illustrate the optimization technique, which can be generalized to any number of states.

In what follows, we find the optimal parameters c^* and $\delta^* = \{t_0^*, t_1^*, t_2^*\}$, and the global minimum through a two step process:

1) For a fixed c, we find the optimal $\delta^c = \{t_0^c, t_1^c, t_2^c\}$ to minimize the objective using the lambda minimization technique.

2) From the objective value obtained in last step, we find an upper bound for c. By enumerating from the minimum base stock level ($c=1$) to the upper bound, we can find the optimal parameters and the global minimum of the objective function.

With c fixed, minimizing (4.43) is equivalent to minimizing

$$v(\delta)=\frac{\phi(c,\delta)-h_S c}{N}=\frac{b_1+KQ(\delta)+b_2 p_L}{M(\delta)}. \tag{4.44}$$

where $b_1=C_1+\frac{(h_W-h_S)}{\mu}, b_2=C_2-C_1-\frac{(h_W-h_S)}{\mu}>0$.

To apply the lambda minimization technique, define the γ-function (analogous to the λ-function in Appendix 4.B) as

$$u(\gamma,\delta)=b_1+b_2 p_L(c,M(\delta))+KQ(\delta)-\gamma M(\delta). \tag{4.45}$$

For a fixed γ, with n = 3 states we have the following optimization problem, where $\delta=\{t_0,t_1,t_2\}$:

$$\begin{aligned}&\min u(\gamma,\delta)\\ &s.t.\ \ t_0\ge t_1\ge t_2\ge 0\end{aligned}$$

Taking partial derivatives of (4.45) with respect to t_0,t_1,t_2 and setting them to 0, we have

$$\frac{\partial u}{\partial t_i}=b_2\frac{\partial p_L}{\partial M}\frac{\partial M}{\partial t_i}+K\frac{\partial Q}{\partial t_i}-\gamma\frac{\partial M}{\partial t_i}=0,\quad i=0,1,2,$$

which is the system of equations that determines the critical point of u. With the partial derivatives of $M(\delta)$ and $Q(\delta)$ developed in Appendix 4.A and equation (4.56), the above system of equations can be reduced to

$$b_2\frac{\partial p_L}{\partial M}+Kh_0(t_0)\Psi(0)-\gamma=0 \tag{4.46}$$

$$b_2\frac{\partial p_L}{\partial M}+Kh_0(t_1)\Psi(1)-\gamma=0 \tag{4.47}$$

$$b_2\frac{\partial p_L}{\partial M}+Kh_0(t_2)\Psi(2)-\gamma=0 \tag{4.48}$$

From (4.46)-(4.48), we have

$$\frac{h_0(t_0)}{h_0(t_1)} = \frac{\Psi(1)}{\Psi(0)} \quad \Rightarrow \quad t_1(t_0) = h_0^{-1}\left[\frac{\Psi(0)}{\Psi(1)} h_0(t_0)\right] \quad \text{and}$$

$$\frac{h_0(t_0)}{h_0(t_2)} = \frac{\Psi(2)}{\Psi(0)} \quad \Rightarrow \quad t_2(t_0) = h_0^{-1}\left[\frac{\Psi(0)}{\Psi(2)} h_0(t_0)\right]$$

Since $h_0(\cdot)$ is monotonically increasing, $t_1(t_0), t_2(t_0)$ are also monotonically increasing functions of t_0. Upon substituting them back into (4.46), we get an univariate equation in t_0, which is

$$b_2 \frac{\partial p_L}{\partial M}(t_0, t_1(t_0), t_2(t_0)) + K h_0(t_0)\Psi(0) - \gamma = 0 \tag{4.49}$$

Note, $\frac{\partial p_L}{\partial M}$ is a function of the tuple $\delta = \{t_0, t_1, t_2\}$, which has been suppressed in its notation.

Lemma 4-3 *For a given γ, the multivariate function $u(\gamma, \delta)$ has a unique critical point.*

Proof.

From Lemma 4-1 and Lemma 4-2, we know that $\frac{\partial p_L}{\partial M}(t_0, t_1(t_0), t_2(t_0))$ is always negative and is a increasing function of t_0. In addition, $h_0(\cdot)$ is an increasing function with $h_0(0) = 0$ and is unbounded as its parameter approaches infinity. Thus the function

$$b_2 \frac{\partial p_L}{\partial M}(t_0, t_1(t_0), t_2(t_0)) + K h_0(t_0)\Psi(0)$$

equals the positive constant γ at a unique point.

Therefore equation (4.49) has a unique solution, which means that the for a given γ, $u(\gamma, \delta)$ has a unique critical point, denoted as $\delta^\gamma = \{t_0^\gamma, t_1^\gamma, t_2^\gamma\}$. ■

The following theorem shows how to find the global minimum of $u(\gamma, \delta)$.

Theorem 4-1 *For a given γ, function $u(\gamma, \delta)$ achieves global minimum at its critical point $\delta^\gamma = \{t_0^\gamma, t_1^\gamma, t_2^\gamma\}$.*

The proof of Theorem 4-1 is in Appendix 4.C.

In light of Theorem 4-1 and the lambda minimization technique, we state the following algorithm to find the optimal δ that minimizes $v(\delta)$ for a given c.

Algorithm 4-1

1. Initialize the iteration counter $m = 0$ and $\gamma = \gamma^0$.
2. For γ^m, use Theorem 4-1 to find $\delta^{c,m} = \{t_0^{c,m}, t_1^{c,m}, t_2^{c,m}\} = \arg\min_{\delta} u(\gamma^m, \delta)$.
3. Use the replacement policy $\delta^{c,m} = \{t_0^{c,m}, t_1^{c,m}, t_2^{c,m}\}$ obtained in step 2 and equation (4.44) to update $\gamma^{m+1} = v(\delta^{c,m})$.
4. If $\gamma^{m+1} = \gamma^m$, stop with $v^c = \gamma^{m+1}$ and $\delta^c = \delta^{c,m}$; otherwise, set $m \leftarrow m+1$ and go to step 2.

In addition, an upper bound on the optimal stock level c can be obtained. Denote the optimal parameters as $\delta^* = \{t_0^*, t_1^*, t_2^*\}$ and c^*. Let $\phi_0 = \phi(c_0, \delta^{c_0})$, which is the optimal cost when c is fixed at c_0. Then we have

$$\phi_0 \geq \phi(c^*, \delta^*) \geq h_S c^* + N \frac{C_1 + (h_W - h_S)\mu^{-1} + KQ(\delta^*)}{M(\delta^*)} \geq h_S c^* + N \frac{C_1 + (h_W - h_S)\mu^{-1} + KQ(\delta')}{M(\delta')}$$

where the second inequality follows by omitting the p_L term in $\phi(c^*, \delta^*)$, and the third inequality holds if δ' is the replacement policy that minimizes the term

$$\frac{C_1 + (h_W - h_S)\mu^{-1} + KQ(\delta)}{M(\delta)}.$$

Using the methods developed in Wu and Ryan (2011), we can obtain δ' as an optimal policy for the condition-based replacement model described there with preventive replacement cost $C_1 + (h_W - h_S)\mu^{-1}$ and additional failure cost K. Thus an upper bound for c^* is given by

$$\overline{c} = \left\lfloor \left(\phi_0 - N \frac{C_1 + (h_W - h_S)\mu^{-1} + KQ(\delta')}{M(\delta')} \right) h_S^{-1} \right\rfloor \tag{4.50}$$

In light of the above discussion, the following algorithm is presented to find the optimal parameters and the global minimum of (4.43).

Algorithm 4-2

1 Initialize $c = c_0$.

2 For fixed c_0, using Algorithm 4-1 to find the optimal parameters $t_i^{c_0}, i = 0,1,2$. Set $\phi_0 = \phi(c^0, t_0^{c_0}, t_1^{c_0}, t_2^{c_0})$.

3 Obtain an upper bound $\bar{c}$ for c^* using equation (4.50).

4 For $c = 1, \cdots, \bar{c}$, find the optimal $t_i^c, i = 0,1,2$ and the corresponding optimal cost $\phi_c^* = \phi(c, t_0^c, t_1^c, t_2^c)$. Then the optimal stock level $c^* = \arg\min_c \phi_c^*$, the optimal replacement policy $\delta^* = \{t_0^*, t_1^*, t_2^*\}$ and the global minimum is $\min \phi_c^*$.

4.5 Numerical Example

Suppose the covariate process $Z(t)$ is a pure birth process with three states $\{0,1,2\}$ and transition rates $v_0 = v_1 = -\ln(0.4), v_2 = 0$. Let $h_0(t) = 2t$ and $\Psi(Z_t) = \exp(2Z_t)$. Assume

$$C_1 = 5, C_2 = 15, K = 25, N = 10, \alpha = 0.2, \beta = 0.5, h = 0.5, \mu = 5.$$

The total cost per unit time is

$$\phi(c, \delta) = 1.5c + 10\frac{4.9 + 10.1 p_L(c, \delta) + 25Q(\delta)}{M(\delta)}.$$

Let $c_0 = 10$. In Algorithm 4-1 step 1, let $\gamma_0 = 12$. As shown in Table 4-1, the lambda-minimization converges after four iterations with $v^{c_0} = 24.7330$ and the optimal parameter $\delta^{c_0} = \{0.5440,\ 0.0736,\ 0.0100\}$. The corresponding cost $\phi_0 = Nv^{c_0} + h_S c_0 = 262.33$.

Table 4-1 Illustration of Algorithm I and the Lambda-Minimization Process

m	γ	Critical point $\delta^{c,m}$	Value of u at critical point	$v(\delta^{c,m})$
0	12	(0.3853, 0.0521, 0.0070)	4.4924	26.4098
1	26.4908	(0.5705, 0.0772, 0.0104)	-0.6750	24.7573
2	24.7573	(0.5444, 0.0737, 0.0100)	-0.0096	24.7330
3	24.7330	(0.5444, 0.0737, 0.0100)	0	24.7330

Based on methods developed in Wu and Ryan (2011), the replacement policy that minimizes term $\frac{4.9+25Q(\delta)}{M(\delta)}$ is $\delta' = \{0.4826, 0.0653, 0.0088\}$ and the minimal value of the term is 24.1302. Thus from equation (4.50), an upper bound for c is

$$\bar{c} = \left\lfloor \left(262.33 - 10\frac{4.9+25Q(\delta')}{M(\delta')} \right) \frac{1}{1.5} \right\rfloor = 14 .$$

In step 4 of Algorithm 4-2, when c ranges from 1 to 14, the resulting total costs are shown in Figure 4-3. The optimal parameters are $c^* = 12$ and $\delta^* = \{0.5048,\ 0.0683,\ 0.0092\}$. The global minimal cost is $\phi^* = 260.827$.

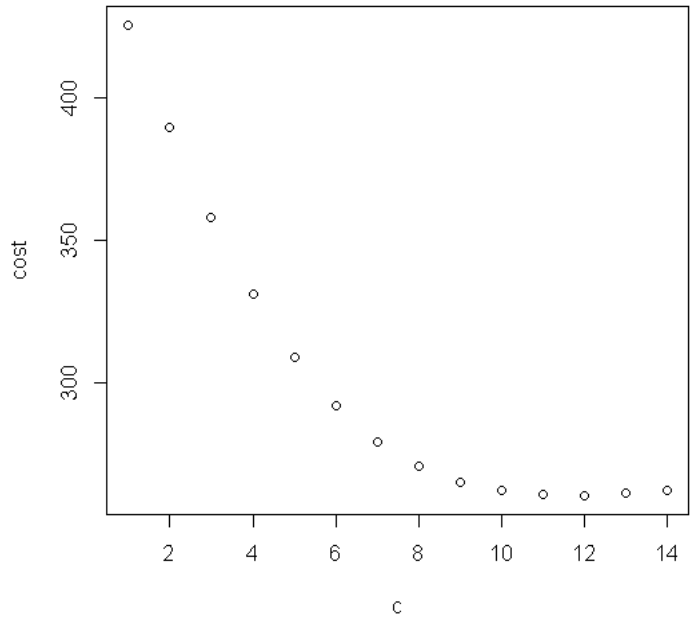

Figure 4-3 The minimized total cost when *c* varies from 1 to its upper bound

4.6 Evaluation of the Single Category Return Assumption

In the previous analysis, we consider the returns as a single category. However, in reality, there is usually a quality difference between the preventively replaced products and failure replaced products in terms of remanufacturing time and remanufacturing cost.

To understand the effect of the single category assumption, in this section, we will examine the case when we categorize the returns into two types: Type 1 (T1), preventively replaced products and Type 2 (T2), failure replaced products, estimate its cost and compare to the cost of the no categorization case.

In general, T1 products have better quality than T2 products, so it typically requires less remanufacturing effort to bring them to the "as good as new" condition. To quantify the quality difference, assume the statement “T1 is x% better than T2 in quality” implies that

1) The unit remanufacturing cost of T1, C_{11}, is x% lower than that of T2, C_{12}; i.e., $C_{11} = C_{12}(1 - x\%)$.
2) The remanufacturing time of T1, $1/\mu_1$, is x% shorter than that of T2, $1/\mu_2$; i.e., $1/\mu_1 = (1 - x\%)/\mu_2$.

4.6.1 Model Analysis

Let $W_1(t), W_2(t)$ be the number of T1, T2 products, respectively, in WIP at time t .Then

$$W_1(t) + W_2(t) + I(t) = c.$$

And let λ_1, λ_2 be the arrival rate of T1, T2 products, respectively. Assume the product mix that enters the remanufacturing process is the same as the product mix that enters the remanufacturable inventory at all times. Then under replacement policy δ,

$$\lambda_1 = \lambda(1 - Q(\delta)),\ \lambda_2 = \lambda Q(\delta)$$

where $\lambda = N / M(\delta)$.

It is not hard to see that $(W_1(t), W_2(t) : t \geq 0)$ consists of a continuous-time Markov chain with a finite state space

$$S = \{(i, j) : i + j \leq c, i = 0,1,\ldots,c,\ j = 0,1,\ldots,c\}$$

A typical portion of the transition diagram among those states is shown in Figure 4-4. Although on the boundaries, one or more of the states depicted in the figure may not exist, the transition rates for the rest are valid. It can be verified that this continuous-time Markov chain is irreducible and ergodic, so it has a limiting distribution.

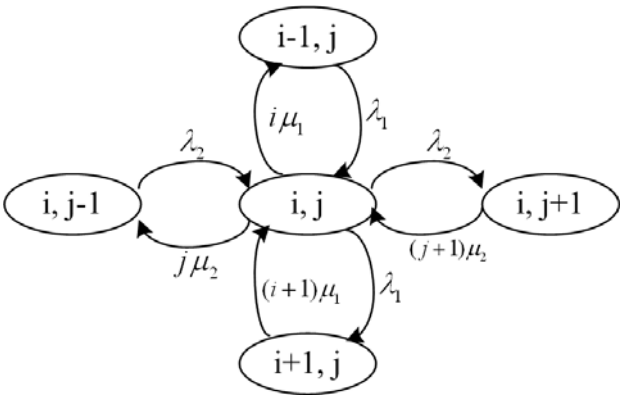

Figure 4-4 Part of transition diagram

Define the indicator function

$$I_S(i,j)=\begin{cases}1 & \text{if } (i,j)\in S\\ 0 & \text{if } (i,j)\notin S\end{cases}$$

Then the balance equations of the limiting probabilities are

$$\begin{aligned}&\lambda_2 P(i,j-1)I_S(i,j-1)+\lambda_1 P(i-1,j)I_S(i-1,j)\\&\quad+(j+1)\mu_2 P(i,j+1)I_S(i,j+1)+(i+1)\mu_1 P(i+1,j)I_S(i+1,j)\\&=[j\mu_2 I_S(i,j-1)+i\mu_1 I_S(i-1,j)\\&\quad+\lambda_2 I_S(i,j+1)+\lambda_1 I_S(i+1,j)]P(i,j)\quad \text{for all } (i,j)\in S\end{aligned}$$

Let $a=\lambda_1/\mu_1, b=\lambda_2/\mu_2$. The solution to the balance equations is

$$P(0,0)=\frac{1}{\sum_{i=0}^{c}\sum_{j=0}^{c-i}\frac{a^i b^j}{i!\,j!}} \tag{4.51}$$

$$P(i,j)=\frac{a^i b^j}{i!\,j!}P(0,0)\quad \text{for all } (i,j)\in S \tag{4.52}$$

Therefore

$$P(W_1(t)+W_2(t)=m)=\sum_{i=0}^{m}P(i,m-i)=\frac{(a+b)^m}{m!}P(0,0)$$

$$p_{L,cat}=P(W_1(t)+W_2(t)=c)=\frac{(a+b)^c}{c!}P(0,0) \tag{4.53}$$

$$E(W)_{cat}=E(W_1(t)+W_2(t))=P(0,0)\sum_{k=0}^{c-1}\frac{(a+b)^{k+1}}{k!} \tag{4.54}$$

With categorization, the replacement cost per unit time for a single product is

$$\phi_{R,cat} = \frac{(1-p_{L,cat})[(1-Q(\delta))(C_{11}+KQ(\delta))+Q(\delta)(C_{22}+KQ(\delta))]+p_{L,cat}(C_2+KQ(\delta))}{M(\delta)}$$
$$= \frac{C_{11}+(K+C_{12}-C_{11})Q(\delta)+p_{L,cat}[C_2-C_{11}-Q(\delta)(C_{12}-C_{11})]}{M(\delta)}$$

and the total cost per unit time is

$$\phi(c,\delta)_{cat} = N\phi_{R,cat} + h_S c + (h_W - h_S)E(W)_{cat}. \quad (4.55)$$

We can follow the two step process as described in section 4.4 to optimize this objective function. However, since this objective function involves more complicated expressions of the loss probability and mean WIP than that of (4.43), minimizing the γ-function in the lambda minimization technique is challenging. Thus for step 1, we resort to some numerical optimization methods to minimize the objective function for a given c. Because first derivatives are available, we adopt the BFGS method (Fletcher, 1987), which is generally considered as the best quasi-Newton method.

The BFGS method cannot guarantee the global optimality of the obtained policy. However, for the single category case, we have verified that the policy and cost obtained using the BFGS method is the same as the optimal policy and cost obtained using lambda minimization technique. Since objective function for the two category case shares the same basic structure with the objection function for single category case, we use the result of BFGS method to approximate the optimal policy and cost in the two category case.

4.6.2 Cost Impact of the Single Category Assumption

Here we illustrate the cost impact of the single category assumption through a numeric example.

Assume T1 products are 20% better than T2 products in quality. Let $C_{11} = 5, \mu_1 = 5$. Then $C_{12} = 6.125, \mu_2 = 4$. Assume all the other parameters stay the same. Then the optimal cost $\phi^*_{cat} = 265.789$ and the optimal parameters are $c^*_{cat} = 12$, $\delta^*_{cat} = \{0.4911, 0.0620, 0.0091\}$.

If the decision maker uses the single category assumption, then first he must estimate the equivalent unit remanufacturing cost

$$C_1 = C_{11}(1-Q(\delta)) + C_{12}Q(\delta)$$

and equivalent unit processing rate

$$\mu = \frac{\lambda}{E(W)}(1-p_L)$$

where $\lambda = N / M(\delta)$. The estimation requires a realization of the operation policy $\{c, \delta\}$. Assume the policy maker can observe the results of $\{c^*_{cat}, \delta^*_{cat}\}$. Then under policy $\{c^*_{cat}, \delta^*_{cat}\}$

$$Q(\delta) = 0.1619,\ M(\delta) = 0.3707,\ p_L = 0.0075,\ E(W) = 5.5710$$

and the estimation would be

$$C_1 = 5.182, \mu = 5.000.$$

With those parameters, the optimal policy under the single category assumption is

$$c_{no\_cat} = 12,\ \delta_{no\_cat} = \{0.5120, 0.0693, 0.0094\}.$$

Under this policy, the actual total cost is $\phi_{no\_cat} = 265.916$, which is 0.05% bigger than δ^*_{cat}. This negligible cost difference indicates that the single category return assumption is acceptable in this case.

Intuitively, the cost difference between the categorized and non-categorized cases depends on the quality difference between T1 and T2. To further evaluate the impact of single category assumption, we vary the quality difference, while keeping $C_{11} = 5, \mu_1 = 5$ unchanged, and then obtain the corresponding cost differences following a similar procedure as above. The results are summarized in Table 4-2.

As expected, for bigger quality difference, the additional cost introduced by the single category assumption is more substantial. And we can see that for our example, as long as the quality difference is below 50%, the cost error caused by the single category assumption is under 1%. Another observation is that as the quality difference increases, we tend to perform the preventive replacement more frequently and keep the stock level higher in the two-category case, which are reasonable because, with a wider quality difference, it is more costly and time consuming to remanufacture a failure-replaced product.

Table 4-2 The Impact of Single Category Assumption under Various Quality Difference between the Two Types of Products

Quality difference x%	c_{cat}^{*}	δ_{cat}^{*}	ϕ_{cat}^{*}	c_{no_cat}	δ_{no_cat}	ϕ_{no_cat}	Cost difference $(\phi_{no\_cat} - \phi_{cat}^{*})/\phi_{cat}^{*}$
20%	12	{0.4911, 0.0620, 0.0091}	265.789	12	{0.5120, 0.0693, 0.0094}	265.916	0.05%
40%	13	{0.4697, 0.0638, 0.0088}	276.206	12	{0.5387, 0.0729, 0.0105}	277.673	0.53%
50%	13	{0.4603, 0.0622, 0.0087}	283.546	12	{0.5323, 0.0720, 0.0102}	285.766	0.78%
60%	14	{0.4364, 0.0591, 0.0081}	294.168	12	{0.5710, 0.0773, 0.0016}	301.289	2.42%
80%	17	{0.3634, 0.0492, 0.0069}	341.987	13	{0.6146, 0.0832, 0.0116}	374.794	9.60%

4.7 Conclusion

This paper investigates a joint operation problem in the context of a product-service system, which to the best of our knowledge has not been addressed in the literature. The system consists of a service subsystem and a remanufacturing subsystem where the replacement decision and the inventory management decision must be made at the same time. Identifying and formulating the couplings between the two subsystems, an integrated model aiming to minimize the total cost per unit time of the system is developed and an algorithm is presented to jointly optimize the replacement policy and the inventory management policy. Then we evaluate the cost impact of treating the preventively replaced products and products replaced due to failure as one category. A numerical example demonstrates that as long as the quality difference between the two types of replaced products is not too large, where how large depends on other parameters in the model, the single category assumption is reasonable.

In this paper, for illustration the covariate process is assumed to have three states, which could be characterized as “like new,” “deteriorated,” and “critical.” It is straightforward to generalize our model to accommodate a finer-grained approximation of a continuous state space by adding more discrete states. The additional effort required for formulation mainly

lies in obtaining the explicit expressions of the mean replacement time and the failure probability for a given replacement policy, which is discussed in detail in Wu and Ryan (2011). Correspondingly, the additional computational effort lies in the evaluation of the mean replacement time and the failure probability. In particular, for an n-state covariate process, the expressions of the mean replacement time and the failure probability consist of several n-fold integrals. Monte Carlo integration methods are essential to evaluate them efficiently (Press et al., 2007).

Other possible extensions to this paper are as follows. First, in our analysis, the demand process of the fleet for new products, which is a superposed renewal process, is approximated by a Poisson process assuming that the number of products in the fleet is sufficiently large. Evaluating the impact of this approximation in the situation of moderate or small fleet sizes is a possible extension of this research. Second, considering the capacity expansion problem of service subsystem in addition to maintenance would be an interesting and challenging problem, which is a natural generalization of the model presented in this study. Third, in a hybrid business model, where the producers operate traditional product sales as well as a PSS, the external returns in addition to the internal replaced products will become part of the input to the remanufacturing system. In this case, the inventory model needs to be reconsidered. Last but not least, the accelerated failure time (AFT) model (Meeker and Escobar, 1998) is considered as a strong competitor to the proportional hazards model when incorporating the covariate information into system failure time estimation. In case it is hard to decide which model to use, the general proportional hazards model (Bagdonavicius and Nikulin, 2001), which includes PHM and the AFT models as special cases, might be appropriate.

Appendix 4.A The Explicit Expressions of $M(\delta)$ and $Q(\delta)$ for PH model with Three-State Covariate Process and Their Partial Derivatives

Assume the covariant process $Z(t)$ has three states $\{0, 1, 2\}$. Let S_i be the product age at which the state changes from i to $i+1$, $i=0,1$. And let $g_0(s_0)$ be the pdf of S_0, $g_1(s_0,s_1)$ be the joint pdf of S_0, S_1. Denote replacement policy $\delta=\{t_0,t_1,t_2\}$. Then the expected length of a replacement cycle $M(\delta)$ and the failure probability $Q(\delta)$ in terms of t_0,t_1,t_2 are

$$M(\delta)=\int_{t_0}^{\infty} M_0(t_0)g_0(s_0)ds_0+\int_{t_1}^{t_0} M_1(s_0)g_0(s_0)ds_0+\int_{t_1}^{\infty}\int_0^{t_1} M_2(s_0,t_1)g_1(s_0,s_1)ds_0ds_1$$
$$+\int_{t_2}^{t_1}\int_0^{s_1} M_3(s_0,s_1)g_1(s_0,s_1)ds_0ds_1+\int_0^{t_2}\int_0^{s_1} M_4(s_0,s_1,t_2)g_1(s_0,s_1)ds_0ds_1,$$

$$Q(\delta)=\int_{t_0}^{\infty} Q_0(t_0)g_0(s_0)ds_0+\int_{t_1}^{t_0} Q_1(s_0)g_0(s_0)ds_0+\int_{t_1}^{\infty}\int_0^{t_1} Q_2(s_0,t_1)g_1(s_0,s_1)ds_0ds_1$$
$$+\int_{t_2}^{t_1}\int_0^{s_1} Q_3(s_0,s_1)g_1(s_0,s_1)ds_0ds_1+\int_0^{t_2}\int_0^{s_1} Q_4(s_0,s_1,t_2)g_1(s_0,s_1)ds_0ds_1$$

where

$$M_0(t_0)=\int_0^{t_0} tdF_0(t)+t_0(1-F_0(t_0))$$

$$M_1(s_0)=\int_0^{s_0} tdF_0(t)+s_0(1-F_0(s_0))$$

$$M_2(s_0,t_1)=\int_0^{s_0} tdF_0(t)+\int_{s_0}^{t_1} tdF_1(s_0,t)+t_1(1-F_1(s_0,t_1))$$

$$M_3(s_0,s_1)=\int_0^{s_0} tdF_0(t)+\int_{s_0}^{s_1} tdF_1(s_0,t)+s_1(1-F_1(s_0,s_1))$$

$$M_4(s_0,s_1,t_2)=\int_0^{s_0} tdF_0(t)+\int_{s_0}^{t_1} tdF_1(s_0,t)+\int_{s_1}^{t_2} tdF_2(s_0,s_1,t)+t_2(1-F_2(s_0,s_1,t_2))$$

$$Q_0(t_0)=F_0(t_0) \qquad Q_1(s_0)=F_0(s_0)$$

$$Q_2(s_0,t_1)=F_1(s_0,t_1) \qquad Q_3(s_0,s_1)=F_1(s_0,s_1)$$

$$Q_4(s_0,s_1,t_2)=F_2(s_0,s_1,t_2)$$

and

$$F_0(t) = 1 - exp\left(-\Psi(0)\int_0^t h_0(u)du\right), t \le s_0$$

$$F_1(s_0,t) = 1 - exp\left(-\Psi(0)\int_0^{s_0} h_0(u)du - \Psi(1)\int_{s_0}^t h_0(u)du\right), s_0 \le t \le s_1$$

$$F_2(s_0,s_1,t) = 1 - exp\left(-\Psi(0)\int_0^{s_0} h_0(u)du - \Psi(1)\int_{s_0}^{s_1} h_0(u)du - \Psi(2)\int_{s_1}^t h_0(u)du\right), s_1 \le t.$$

The partial derivatives of $M(\delta)$ and $Q(\delta)$ with respect to t_i are

$$\frac{\partial M(\delta)}{\partial t_0} = \left(1 - \int_0^{t_0} g_0(s_0)ds_0\right)(1 - F_0(t_0))$$

$$\frac{\partial M(\delta)}{\partial t_1} = \int_{t_1}^{\infty}\int_0^{t_1}(1 - F_1(s_0,t_1))g_1(s_0,s_1)ds_0ds_1$$

$$\frac{\partial M(\delta)}{\partial t_2} = \int_0^{t_2}\int_0^{s_1}(1 - F_2(s_0,s_1,t_2))g_1(s_0,s_1)ds_0ds_1$$

$$\frac{\partial Q(\delta)}{\partial t_0} = h_0(t_0)\Psi(0)\left(1 - \int_0^{t_0} g_0(s_0)ds_0\right)(1 - F_0(t_0))$$

$$\frac{\partial Q(\delta)}{\partial t_1} = h_0(t_1)\Psi(1)\int_{t_1}^{\infty}\int_0^{t_1}(1 - F_1(s_0,t_1))g_1(s_0,s_1)ds_0ds_1$$

$$\frac{\partial Q(\delta)}{\partial t_2} = h_0(t_2)\Psi(2)\int_0^{t_2}\int_0^{s_1}(1 - F_2(s_0,s_1,t_2))g_1(s_0,s_1)ds_0ds_1$$

Note that

$$\frac{\partial Q(\delta)}{\partial t_i} = h_0(t_i)\Psi(i)\frac{\partial M(\delta)}{\partial t_i}, \forall i. \tag{4.56}$$

Appendix 4.B Lambda Minimization Technique

In this section, we will give a brief introduction to the lambda minimization technique developed in Aven and Bergman (1986), aiming to minimizing a function with the following form

$$B(X) = \frac{M(X)}{S(X)} \tag{4.57}$$

where $X \in \Re^n$ and $S(X) > 0, \forall X$.

Define the λ-function

$$C(X,\lambda) = M(X) - \lambda S(X) \qquad \lambda \in (-\infty, +\infty).$$

For each λ, denote the value of X that minimizes $C(X,\lambda)$ as X_λ.

Now we will show how the problem of minimizing $B(X)$ can be solved by minimizing the λ-function $C(X,\lambda)$. Aven and Bergman proved the following proposition which associates the optimality of $B(X)$ with the optimality of $C(X,\lambda)$.

Proposition 4-1: *If X_λ minimizes $C(X,\lambda)$ and $C(X_\lambda,\lambda)=0$, then X_λ is optimal for* (4.57) *and the optimal value of $B(X)$ is $B(X_\lambda)=\lambda\equiv\lambda^*$.*

Aven and Bergman then proved another important proposition, stated below, which leads to an iteration algorithm that always produces a sequence converging to λ^*.

Proposition 4-2: *Choose any λ_1 and set iteratively $\lambda_{n+1}=B(X_{\lambda_n})$, $n=1,2,3,\cdots$. Then*

$$\lim_{n\to\infty}\lambda_n = \lambda^*.$$

Propositions 1 and 2 imply that the minimization of $B(X)$ can be transformed into the problem of minimizing $C(X,\lambda)$ plus a succession of iterations. This is the essence of the lambda minimization technique. This technique is very suitable for situations where it is easy to find the optimal solutions to the λ-function $C(X,\lambda)$ while it is hard to minimize $B(X)$ directly; this is often the case in replacement/maintenance applications. The optimal solution and the optimal value of $B(X)$ can be attained simultaneously when the algorithm converges.

Appendix 4.C Proof of Theorem 4-1

For readability, first we list all the monotonicity properties of various functions that are related to the proof of Theorem 4-1 in the following.

1) $M(\delta)$ is increasing in t_i, $\forall i$.
2) $\dfrac{\partial p_L}{\partial M}$ is increasing in t_i, $\forall i$.
3) $h_0(\cdot)$ is increasing in its parameter.

4) t_1 is increasing in t_0 if $t_1 = h_0^{-1}\left[\frac{\Psi(0)}{\Psi(1)} h_0(t_0)\right]$ and t_2 is increasing in t_0 if $t_2 = h_0^{-1}\left[\frac{\Psi(0)}{\Psi(2)} h_0(t_0)\right]$.

Since γ is fixed, in the following discussion, it is suppressed in the notation of u.

The feasible region of $u(\delta)$ is $R = \{\delta : t_0 \geq t_1 \geq t_2 \geq 0\}$. Divide this region into two sets: a closed and bounded set $D = \{\delta : \Lambda \geq t_0 \geq t_1 \geq t_2 \geq 0\}$ where Λ is an arbitrary large positive number, and set $B = R \setminus D$; i.e., $\{B, D\}$ is a partition of R. Define $t_i = +\infty$ represent failure replacement in state *i*.

Lemma 4-4 *$u(\delta)$ achieves its minimum at the critical point $\delta^\gamma = \{t_0^\gamma, t_1^\gamma, t_2^\gamma\}$ in D.*

Proof

Since u is a continuous function on the closed and bounded region D, according to the extreme value theorem for multivariate functions (Stewart, 1999), u has a global minimum which happens either at its critical point or a certain point on the boundary.

The boundaries of $u(\delta)$ are where $t_0 = t_1$, $t_1 = t_2$, $t_2 = 0$ or $t_0 = \Lambda$ within the feasible region $D = \{\delta : \Lambda \geq t_0 \geq t_1 \geq t_2 \geq 0\}$. We prove by contradiction that points on the boundaries are not optimal to $u(\delta)$.

1) Points on the boundary where $t_0 = t_1$.

Assume an optimal point exists on this boundary and denote it as $\delta = \{a_0, a_0, a_2\}$, $a_0 \geq a_2 \geq 0$. Recall that the partial derivative of $u(\delta)$ with respect to t_0, t_1, t_2 are

$$\frac{\partial u}{\partial t_i} = \frac{\partial M}{\partial t_i}\left(b_2 \frac{\partial p_L}{\partial M}(t_0, t_1, t_2) + K h_0(t_i)\Psi(i) - \gamma\right), \quad i = 0, 1, 2,$$

and $\frac{\partial M}{\partial t_i} > 0, \forall i$.

i. If $b_2 \frac{\partial p_L}{\partial M}(a_0, a_0, a_2) + K h_0(a_0)\Psi(0) - \gamma \geq 0$

it follows that

$$b_2 \frac{\partial p_L}{\partial M}(a_0, a_0, a_2) + Kh_0(a_0)\Psi(1) - \gamma > 0 .$$

According to definition of a continuous function,

$$\exists a_1 \in (a_2, a_0) \text{, s.t. } b_2 \frac{\partial p_L}{\partial M}(a_0, a_1, a_2) + Kh_0(a_1)\Psi(1) - \gamma > 0 .$$

From assumption #7 and Lemma 4-2, we know that $b_2 \frac{\partial p_L}{\partial M}(t_0, t_1, t_2) + Kh_0(t_1)\Psi(i)$ is increasing in t_1. Thus

$$\frac{\partial u}{\partial t_1}(a_0, t_1, a_2) > 0 \ \forall t_1 \in [a_1, a_0],$$

which means u is an increasing function for $t_1 \in [a_1, a_0]$ with fixed $t_0 = a_0, t_2 = a_2$. Therefore $u(\{a_0, a_1, a_2\}) < u(\{a_0, a_0, a_2\})$. Assumption disproved.

ii. If $b_2 \frac{\partial p_L}{\partial M}(a_0, a_0, a_2) + Kh_0(a_0)\Psi(0) - \gamma < 0$

It follows that

$$\exists a_0' > a_0 \text{, s.t. } b_2 \frac{\partial p_L}{\partial M}(a_0', a_0, a_2) + Kh_0(a_0')\Psi(0) - \gamma < 0 .$$

Thus $\frac{\partial u}{\partial t_0}(t_0, a_0, a_2) < 0$, $\forall t_0 \in [a_0, a_0']$, which means u is an decreasing function for $t_0 \in [a_0, a_0']$ with fixed $t_1 = a_0, t_2 = a_2$. Therefore $u(\{a_0', a_0, a_2\}) < u(\{a_0, a_0, a_2\})$. Assumption disproved.

2) Points on the boundary where $t_1 = t_2$

Using the same argument as in 1), we can prove that points on this boundary are not optimal.

3) Points on the boundary where $t_2 = 0$

Assume $(a_0, a_1, 0)$ is optimal.

Since $\frac{\partial p_L}{\partial M} < 0$, $h_0(0) = 0$ and $\gamma > 0$, $b_2 \frac{\partial p_L}{\partial M}(a_0, a_1, 0) + Kh_0(0)\Psi(2) - \gamma < 0$

It follows

$$\exists a_2 \in (0, a_1), \text{ s.t. } b_2 \frac{\partial p_L}{\partial M}(a_0, a_1, a_2) + Kh_0(a_2)\Psi(2) - \gamma < 0.$$

Thus $\frac{\partial u}{\partial t_2}(a_0, a_1, t_2) < 0$, $\forall t_2 \in [0, a_2]$, which mean u is an decreasing function for $t_2 \in [0, a_2]$ with fixed $t_0 = a_0, t_1 = a_1$. Therefore $u(\{a_0, a_1, a_2\}) < u(\{a_0, a_1, 0\})$. Assumption disproved.

4) Points on the boundary where $t_0 = \Lambda$

Since K is arbitrary large and $h_0(\cdot)$ is increasing and unbounded, then we have

$$b_2 \frac{\partial p_L}{\partial M}(\Lambda, a_1, a_2) + Kh_0(\Lambda)\Psi(0) - \gamma > 0$$

It follows

$$\exists a_0 \in [a_1, \Lambda), \text{ s.t. } b_2 \frac{\partial p_L}{\partial M}(a_0, a_1, a_2) + Kh_0(a_0)\Psi(0) - \gamma > 0.$$

Thus $\frac{\partial u}{\partial t_0}(t_0, a_1, a_2) > 0 \ \forall t_0 \in [a_0, \Lambda)$, which means u is an increasing function for $t_0 \in [a_0, \Lambda)$ with fixed $t_1 = a_1, t_2 = a_2$. Therefore $u(\{a_0, a_1, a_2\}) < u(\{\Lambda, a_1, a_2\})$. Assumption disproved.

In summary, in region D, $u(\delta)$ can not achieve its global minimum at its boundaries; which means $u(\delta)$ can only achieve its global minimum at its unique critical point $\delta^\gamma = \{t_0^\gamma, t_1^\gamma, t_2^\gamma\}$. ■

Lemma 4-5 *$\forall \delta \in B, \exists \delta' \in D$, such that $u(\delta') < u(\delta)$.*

Proof

There are only three types of points in B : $\delta = \{\infty, a_1, a_2\}$, $\delta = \{\infty, \infty, a_2\}$ and $\delta = \{\infty, \infty, \infty\}$, $\infty > a_1 > a_2 \geq 0$.

1) For points in the form of $\delta = \{\infty, a_1, a_2\}$

Since $h_0(\cdot)$ is increasing and unbounded, then we have

$$\lim_{t_0 \to \infty} b_2 \frac{\partial p_L}{\partial M}(t_0, a_1, a_2) + Kh_0(t_0)\Psi(0) - \gamma > 0.$$

Thus

$$\exists a_0 \in [a_1, \infty), \text{ s.t. } b_2 \frac{\partial p_L}{\partial M}(a_0, a_1, a_2) + Kh_0(a_0)\Psi(0) - \gamma > 0 .$$

Thus $\frac{\partial u}{\partial t_0}(t_0, a_1, a_2) > 0 \ \forall t_0 \geq a_0$, which means u is an increasing function for $t_0 \geq a_0$ with fixed $t_1 = a_1, t_2 = a_2$. Therefore if we let $\delta' = \{a_0, a_1, a_2\}$, then $\delta' \in D$ and $u(\delta') < u(\delta)$.

2) For points in the form of $\delta = \{\infty, \infty, a_2\}$

Using the similar argument as in 1), there exists $a_1 \in [a_2, \infty)$, such that $\frac{\partial u}{\partial t_1}(\infty, a_1, a_2) > 0 \ \forall t_1 \geq a_1$. Therefore $u(\{\infty, a_1, a_2\}) < u(\delta)$. And based on the result in 1), $\exists \delta' \in D$, such that $u(\delta') < u(\{\infty, a_1, a_2\}) < u(\delta)$.

3) For points in the form of $\delta = \{\infty, \infty, \infty\}$

Similarly, we can prove that

$$\exists \delta' \in D, a_2 \in [0, \infty), \text{ such that } u(\delta') < u(\{\infty, \infty, a_2\}) < u(\delta) .$$

To sum up, $\forall \delta \in B, \exists \delta' \in D$, such that $u(\delta') < u(\delta)$. ■

From Lemma C.1 and Lemma C.2, we conclude that for a given γ, function $u(\gamma, \delta)$ achieves a global minimum at its unique critical point, $\delta^\gamma = \{t_0^\gamma, t_1^\gamma, t_2^\gamma\}$.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant CNS-0540293.

References

Aras, N., Boyaci, T., and Verter, V. (2004). The effect of categorizing returned products in remanufacturing. *IIE Transactions*, 36(4):319–331.

Armstrong, M. J. and Atkins, D. R. (1996). Joint optimization of maintenance and inventory policies for a simple system. *IIE Transactions*, 28:415–424.

Armstrong, M. J. and Atkins, D. R. (1998). A note on joint optimization of maintenance and inventory. *IIE Transactions*, 30:143–149.

Aven, T. and Bergman, B. (1986). Optimal replacement times – a general set-up. *Journal of Applied Probability*, 23:432–442.

Bagdonavicius, V. and Nikulin, M. (2001). *Accelerated life models: modeling and statistical analysis*. Chapman and Hall/CRC.

Baines, T. S. and Lightfoot, H. W. (2007). State-of-the-art in product-service systems. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 221:1543–1552.

Banjevic, D., Jardine, A. K. S., Makis, V., and Ennis, M. (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR*, 39:32–50.

Brezavscek, A. and Hudoklin, A. (2003). Joint optimization of block-replacement and periodic-review spare-provisioning policy. *IEEE Transactions on Reliability*, 52:112 – 117.

Buzacott, J. A. and Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*. Prentice-hall, Inc., Englewood Cliffs, NJ.

Cinlar, E. and Lewis, P. A. W. (1972). Superposition of point processes. In *Stochastic Point Processes*, pages 546–606. Wiley-Interscience, New York.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Das, T. K. and Sarkar, S. (1999). Optimal preventive maintenance in a production inventory system. *IIE Transactions*, 31:537–551.

Fleischmann, M., Bloemhof-Ruwaard, J. M., Dekker, R., Van der Laan, E., and Van Numen, J. A. E. E. (1997). Quantitative models for reverse logistics: A review. *European Journal of Operational Research*, 103:1–17.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons Ltd, 2nd edition.

Ghasemi, S., Yacout, S., and Ouali, M. S. (2007). Optimal condition based maintenance with imperfect information and the proportional hazards model. *International Journal of Production Research*, 45(4):989–1012.

Giorgio, M., Guida, M., and Pulcini, G. (2011). An age- and state-dependent Markov model for degradation processes. *IIE Transactions*, 43:621–632.

Goedkoop, M., van Halen, C., and te Riele, H. (1999). Product service-systems, ecological and economic basics. Report for Dutch ministries of environment (VROM) and economic affairs (EZ). http://www.pre.nl/download/ProductService.zip.

Guide, V. D. R. (2000). Production planning and control for remanufacturing: Industry practice and research needs. *Journal of Operations Management*, 18:467–483.

Guide, V. D. R. and Van Wassenhove, L. N. (2009). The evolution of closed-loop supply chain research. *Operations Research*, 57(1):10–18.

Harel, A. (1990). Convexity properties of the Erlang loss formula. *Operations Research*, 38:499–505.

Ilgin, M. A. and Tunali, S. (2007). Joint optimization of spare parts inventory and maintenance policies using genetic algorithms. *The International Journal of Advanced Manufacturing Technology*, 34:594–604.

Kharoufeh, J. P., Solo, C. J., and Ulukus, M. Y. (2010). Semi-Markov models for degradation-based reliability. *IIE Transactions*, 42:599–612.

Lam, C. T. and Yeh, R. H. (1994). Optimal maintenance policies for deteriorating systems under various maintenance strategies. *IEEE Transactions on Reliability*, 43(3):423–430.

Liu, Y., Li, Y., Huang, H.-Z., Zuo, M. J., and Sun, Z. (2010). Optimal preventive maintenance policy under fuzzy bayesian reliability assessment environments. *IIE Transactions*, 42:734–745.

Luiten, H., Knot, M., and van der Horst, T. (2001). Sustainable product-service-systems: the kathalys method. In *Proceedings of the Second International Symposium on Environmentally conscious design and inverse manufacturing*, pages 190–197.

Manzini, E., Vezzoli, C., and Clark, G. (2001). Product service-systems: using an existing concept as a new approach to sustainability. *Journal of Design Research*, 1(2).

Maxwell, I. and van der Vorst, R. (2003). Developing sustainable products and services. *Journal of Clearner Production*, pages 883–895.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and P., F. B. (2007). *Numerical recipes: the art of scientific computing*. Cambridge University Press, third edition.

Rezg, N., Xie, X., and Mati, Y. (2004). Joint optimization of preventive maintenance and inventory control in a production line using simulation. *International Journal of Production Research*, 42:2029–2046.

Ross, S. M. (2003). *Introduction to Probability Models*. Academic Press, San Diego, CA, 8th edition.

Stewart, J. (1999). *Calculus*. Brooks/Cole Pub Co, 4th edition.

Thierry, M. C., Salomon, M., and Van Nunen, J. A. E. E. (1995). Strategic production and operations management issues in product recovery management. *California Management Review*, 37:114–135.

Van der Laan, E., Salomon, M., Dekker, R., and Wassenhove, L. V. (1999). Inventory control in hybrid systems with remanufacturing. *Management Science*, 45(5):733–747.

Van Halen, C., Vezzoli, C., and Wimmer, R. (2004). *MEPSS Handbook*. Royal Van Gorcum, Assen, Netherlands.

White, A., Stoughton, M., and Feng, L. (1999). *Servicizing: The Quiet Transition to Extended Producer Responsibility*. Tellus Institute, Boston.

Wu, X. and Ryan, S. M. (2010). Value of condition monitoring for optimal replacement in the proportional hazards model with continuous time degradation. *IIE Transactions*, 42:553–563.

Wu, X. and Ryan, S. M. (2011). Optimal replacement in the proportional hazards model with semi-Markovian covariate process and continuous monitoring. *IEEE Transactions on Reliability*, 60:580–589.

CHAPTER 5 GENERAL CONCLUSION

In this dissertation, we studied the condition-based replacement problem for general deteriorating systems whose aging and deterioration process is assumed to follow the proportional hazards (PH) model. The condition information of the system is characterized by a stochastic covariate process. For various covariate processes and various monitoring schemes, we identified the forms of the optimal replacement policies and developed procedures to obtain the optimal policy parameters and optimal costs. In addition, an application of the PH-based models to a product service system was carefully investigated.

In Chapter 2, we considered the condition-based replacement problem for systems in the PH model with continuous time Markov covariate process and periodic monitoring. We followed the model of Makis and Jardine (1992) but removed their discrete-time approximation by explicitly accounting for the possibility that the concomitant Markov chain may make transitions among its states between observation epochs. Accounting for state transitions between observations introduces significant intricacies in the computation of policy parameters. We used conditioning to develop a new recursive procedure to obtain the parameters of the optimal replacement policy and its long-run average cost. Based on that, we compared the costs of three monitoring schemes: no monitoring which corresponds to age-based replacement, periodic monitoring at various intervals, and continuous monitoring (approximated as the interval vanishes). We illustrated the relationships between the unit cost of periodic monitoring and the upfront cost of continuous monitoring under which the continuous, periodic or no monitoring scheme is optimal.

In Chapter 3, we extended the PH-based replacement models to systems with semi-Markov covariate processes and continuous monitoring. We identified our model as a special case of the one described in Bergman (1978), and showed that, if the hazard function of the system is non-decreasing, then the optimal replacement policy of our model is of the control limit type with respect to the hazard function. Given that an optimal policy may be uniquely defined by a set of state-dependent threshold ages for replacement, an explicit expression for

the objective function was derived in terms of those threshold ages by conditioning. Then the iterative procedure developed by Bergman was customized for our model to find the optimal threshold ages. The model and the computation procedure were illustrated by numerical examples, and its computational advantage over the recursive procedure in Wu and Ryan (2010) was discussed. The effect of different sojourn time distributions of the covariate process on the optimal policy and cost was also studied.

In Chapter 4, we investigated a joint operation problem in the context of a product-service system. The system consists of a service subsystem and a remanufacturing subsystem where the condition-based replacement decision and the inventory management decision must be made at the same time. Identifying and formulating the couplings between the two subsystems, an integrated model aiming to minimize the total cost per unit time of the system was developed and an algorithm was presented to jointly optimize the replacement policy and the inventory management policy. Then we evaluated the cost impact of treating as one category the preventively replaced products and products replaced due to failure.

Future research directions for the PH-based replacement models could be

- Generalize the one-dimensional covariate to a multi-dimensional vector which would permit the covariate process to evolve along multiple paths.
- Introduce uncertainty in the monitoring process. Extend our models to systems with imperfect monitoring; that is, where the information obtained through monitoring can only be used to calculate the probability that the system is in a certain diagnostic state.

In the analysis of the joint optimization of the PSS, the demand process of the fleet for new products, which is a superposed renewal process, is approximated by a Poisson process assuming that the number of products in the fleet is sufficiently large. Evaluating the impact of this approximation in the situation of moderate or small fleet sizes is a possible extension of this research. Also, considering the capacity expansion problem of the service subsystem in addition to replacement would be interesting and challenging problem, which is a natural generalization of the model presented in Chapter 4.

References

Bergman, B. (1978). Optimal replacement under a general failure model. Adv. in Appl. Prob.,10(2):431–541.

Makis, V. and Jardine, A. K. S. (1992). Optimal replacement in the proportional hazards model. INFOR, 30(1):172–183.

Wu, X. and Ryan, S. M. (2010). Value of condition monitoring for optimal replacement in the proportional hazards model with continuous time degradation. IIE Transactions, 42:553–563.